PK
555
[illegible]
[illegible]

ENDEAVOUR VOL 23 1964 DR. S.R. BOSE.

077815

गुरुकुल कांगड़ी विश्वविद्यालय, हरिद्वार
पुस्तकालय

विषय संख्या R 505
पुस्तक संख्या L84E V.23
आगत पंजिका संख्या ~~65445~~

पुस्तक पर किसी प्रकार का निशान लगाना वर्जित है। कृपया १५ दिन से अधिक समय तक पुस्तक अपने पास न रखें।

पुस्तकालय, गुरुकुल कांगड़ी विश्वविद्यालय
हरिद्वार।

~~63443~~

COMPILED

R
505
L84E
V. 25

पुस्तकालय

गुरुकुल कांगड़ी विश्वविद्यालय, हरिद्वार

पुस्तक-वितरण की तिथि नीचे अंकित है।

इस तिथि सहित १५वें दिन तक यह पुस्तक पुस्तकालय में वापिस आ जानी चाहिए। अन्यथा ५ पैसे प्रतिदिन के हिसाब से विलम्ब-दण्ड लगेगा।

077815

STOCK VERIFICATION
Central Library
2024-25 G.K.D.V., Hdr.

स्टॉक प्रमाणीकरण १९८४-१९८५

63443

88

ENDEAVOUR

VOLUME XXIII JANUARY 1964

COMPILED

A review of the progress of science, published in five languages

पुस्तक सं. ....
आगत नं. ....
दिनांक ....

CHECKED 1973
Initial

Editor TREVOR I. WILLIAMS, M.A., B.Sc., D.PHIL., F.R.I.C.
Foreign Editor G. N. J. BECK, B.Sc., A.M.I.E.E.
Assistant Editor W. A. OSMAN, B.Sc., M.Sc., A.R.C.S.

Contents

077815

92:51/53
2 Editorial: Galileo Galilei (1564–1642)

546.29
3 The chemistry of the noble gases
N. Bartlett, B.Sc., Ph.D.

548.526:539.4
8 Whiskers
J. E. Gordon, B.Sc.

007:621.391
13 The communication of information
Colin Cherry, D.Sc. (Eng.), A.M.I.E.E., Hon. A.C.G.I.

92:521
18 William Herschel and sidereal astronomy
Michael Hoskin, Ph.D.

541.11
22 Modern thermochemistry
C. T. Mortimer, M.Sc., Ph.D., F.R.I.C., and
H. D. Springall, M.A., D.Phil., F.R.I.C.

551.509.3
27 Weather forecasting by electronic calculation
R. C. Sutcliffe, C.B., O.B.E., Ph.D., F.R.S.

576.314:576.85:547.964.4
33 Teichoic acids and the bacterial cell wall
J. Baddiley, Ph.D., D.Sc., F.R.S.

591.53:595.772.4
38 Feeding habits of tsetse flies
B. Weitz, D.Sc., M.R.C.V.S.

551.46(267)
43 Geophysical studies in the Indian Ocean
B. D. Loncarevic, M.A., Ph.D.

47 Book reviews

Some books received (cover iii)

Communications to the Editor should be addressed to South Block, Thames House, Millbank, London S.W.1. Telephone: Tate Gallery 2181

Galileo Galilei (1564–1642)

Galileo Galilei was born at Pisa four hundred years ago, on 15th February 1564, the year in which Michelangelo died. He died in 1642, the year in which Newton was born. Indisputably, he was one of the truly great men of science: apart from the intrinsic value of his work, especially in the field of dynamics, he had a powerful influence in steering thought about natural phenomena into a new and immensely fruitful channel. Galileo began as an Aristotelian but ended as an ardent advocate of the experimental method, of which he was himself a master. Declaring frankly that he knew nothing of such things as the cause of gravity or the nature of force, he directed all his attention in later life to establishing how things happen. His personal, as opposed to his scientific, greatness is less easily assessed, and has excited much controversy. To some, he is one of the martyrs of science, recanting his beliefs only in the face of the direst threats. Others see his recantation as a lamentable weakness and maintain that his punishment was trivial by the standards of the day.

Galileo was the son of Vincenzo Galilei (1520–91), a musician of some distinction; the son inherited his father's love of music and poetry. His father's intention was that he should become a cloth-dealer, but eventually he became a medical student at Pisa. Here, in 1581, timing the swing of one of the candelabra in the cathedral with his pulse, he discovered that the time of oscillation of a pendulum is independent of its amplitude.

In 1582 the Grand Duke of Tuscany came to Pisa, and his court included a mathematician whose lessons Galileo was able to attend. Although financial difficulties made it impossible for him to remain at the university, he pursued his mathematical studies at home to such good effect that in 1589 he was appointed professor of mathematics at Pisa. He is popularly supposed to have carried out, from the leaning tower of Pisa, experiments refuting the Aristotelian belief that bodies fall at a rate proportional to their weight. There is, however, doubt whether he ever carried out the demonstration popularly ascribed to him. His actual experiment appears to have been to drop a lead and a wooden sphere simultaneously; he observed that although at first the wood appeared to fall more quickly, the lead reached the ground first. Galileo's colleagues were hostile to him, and in consequence he resigned his chair and retired to Florence. In 1592, he became professor of mathematics at Padua, where he remained until 1610. He established a reputation as a lecturer, but his new colleagues proved to have little more respect for experimental evidence than his old ones. There seems little doubt, however, that Galileo's irascibility and sarcastic tongue contributed much to the troubles that beset his life.

Although the early history of the telescope is obscure, it is certain that Galileo was among the first to construct an effective instrument, having heard of the principle from Holland, where the first examples were constructed. With this, in January 1610, he discovered the satellites of Jupiter, and later in the same year he observed the phases of Venus. He also observed the lunar landscape, and discovered that the apparently smooth band of the Milky Way consisted, in fact, of a vast multitude of stars. He correctly identified sunspots as markings on the solar surface and from their movement deduced the rate of rotation of the Sun.

To the Padua period belong other important practical discoveries. His air thermometer dates from 1602, and is a precursor of modern instruments. He was interested also in magnetism and in practical mechanics, establishing a workshop in his own house for making instruments—including many telescopes—that found their way all over Europe. But it is, of course, for his contribution to dynamics that he earned lasting fame. To these we now turn, but to see them in proper perspective some brief account of earlier work must be given.

Although heliocentric theories of the Universe had been advanced from classical times—notably by Aristarchus, though Pythagoras, too, had dared to ascribe a movement to the Earth—they had gained little acceptance among western scholars. Copernicus (1473–1543) was the first to advance a theory of the solar system that was in agreement with general observation. His views were set forth about 1530 in his *De Revolutionibus Orbium Coelestium*, which, thanks to a tactful preface by Osiander, earned the approval of Pope Clement VII, who asked that it should be published in full. The favourable papal attitude to the Copernican theory was not lasting. In 1616, because Galileo's work had revived interest in the subject in Italy, strong ecclesiastical opposition arose and Copernicus's book was proscribed, though it was released again in 1620 in an only slightly revised form. Johann Kepler (1571–1630) gave greater precision to the Copernican theory, establishing the three laws of planetary motion.

Galileo, testing his assumptions with balls running down inclined planes, to get over the difficulty resulting from the relatively high speeds of freely falling bodies, established the law of motion for such bodies. His investigations showed that, contrary to the then general belief, no force is needed to maintain a body in continuous rectilinear motion. This was of immediate significance to astronomical theory: force is not required to keep the planets in motion, but only to deflect them constantly from a straight line so that they travel in orbits round the Sun. But Galileo had no conception of universal gravitation, and it remained for Newton, in 1685, to define this force and the laws governing its action.

Galileo's observation of the phases of Venus gave direct experimental support to the Copernican theory and, despite the ridicule of his colleagues, brought him widespread fame. His *Siderius nuncius* of 1610 was an exposition of the Copernican theory. In 1611, he was received by Pope Paul V, apparently in friendly fashion. Yet he evidently left enemies behind him in Rome; in 1615 he was summoned from Pisa, to which he had returned in 1610, to face accusations of heresy. There is a good deal of evidence that he brought this on himself by obliging the authorities to pronounce on a matter that they had no great desire to pursue. He was compelled to abjure his expressed belief that the Earth moves round the Sun, but allowed to return to live in Florence. In spite of this, he published in 1632 his *Dialogo di Galileo Galilei Linceo*, which took the form of an imaginary argument, a form popular in those days, between advocates of Ptolemaic and Copernican systems. This brought more serious trouble upon him. He was sentenced to imprisonment and obliged to do public penance. At the end of his ordeal he is said to have muttered '*Eppur si muove*' ('but yet it moves'), but the evidence for this is doubtful; the first account of it did not appear until 1761.

He was allowed to begin to serve his sentence in a house belonging to his old patron the Grand Duke of Tuscany. Later he was allowed to move to Siena, and finally to a house of his own at Arcetri. In 1638, his *Discorsi e dimostrazioni matematiche intorno a due nuove scienze attenenti alla meccanica & i movimenti locali* was published at Leyden. This was a comprehensive exposition of his theories of dynamics, and his last major work.

Although Galileo endured humiliations at the hands of some of his compatriots, his reputation stood high elsewhere. John Milton, who makes several allusions to him in his poetry, visited him in 1638 and assured him that his *Dialogo* was being widely and appreciatively read in England. Collected editions of his works appeared long after his death. Now, when the controversies of his age have subsided, his greatness is universally recognized.

The chemistry of the noble gases

N. Bartlett

The conclusive establishment, in 1962, of the existence of compounds of the noble gases has overthrown a long-established chemical belief. This article reviews the preparation and chemistry of these compounds. The molecular structures of the simple xenon compounds are seen to correspond to those of iodine compounds having similar stoichiometries and electronic configurations. The available experimental evidence indicates that the nature of the bonding is similar to that found in other, less exotic, compounds. The article concludes with an assessment of the range of compounds the various noble gases can be expected to form.

The first true compounds of the noble gases were prepared in 1962 [1–3], more than sixty years after the discovery of the gases themselves by Sir William Ramsay. Although many claims to have made compounds in which noble gas atoms were chemically bonded to other atoms had appeared earlier, none had been conclusively established. Clathrate compounds of the noble gases had, however, been known for a number of years; indeed, P. Villard prepared the first, a hydrate of argon, shortly after the discovery of the gas itself [4]. However, there is in such compounds no chemical bonding of the noble gas, the molecules of which are simply caged by the host lattice.

Apparently Moissan, a friend of Ramsay, tried without success to bring about a reaction between argon and fluorine [5]. In his later years, Ramsay was evidently convinced of the chemical inertness of the gases, because he rejected a suggestion by G. Oddo [6] that krypton and xenon ought to form halides. Similar predictions were made by A. von Antropoff [7], who went so far as to predict a possible octavalence for the heavier gases; he and his co-workers attempted to make compounds, but they had no unequivocal success. In 1933, L. Pauling [8] saw that the heavier gases were the more likely to form compounds, and that fluorine and oxygen were the elements most likely to enter into combination with them. He suggested that hexafluorides of krypton and xenon, and an octafluoride of the latter, might be made. In the same prophetic communication he also predicted the existence of an acid H_4XeO_6 and salts of octavalent xenon, such as Ag_4XeO_6. In the same year, unsuccessful attempts were made to synthesize fluorides of argon and krypton by O. Ruff and W. Menzel [9] and of xenon by D. M. Yost and A. L. Kaye [10]. The latter workers, experimenting with electrical discharges through mixtures of xenon and fluorine, came closest to success. Unfortunately, they conducted their experiments in quartz apparatus. They concluded that the etching of their apparatus might have been due to the action of a xenon fluoride, and that the experiments did not rule out the existence of such a compound. Had their apparatus been efficiently cooled, or constructed of fluorite or alumina instead of quartz, the xenon fluorides might well have been discovered thirty years ago. In the event, the failure to confirm the existence of any of the predicted compounds, and the success of the early electronic theories of valence, which stressed the stability of the inert-gas configuration, combined to entrench the conclusion that the noble gases were incapable of chemical reaction.

Preparation of the first true noble-gas compounds

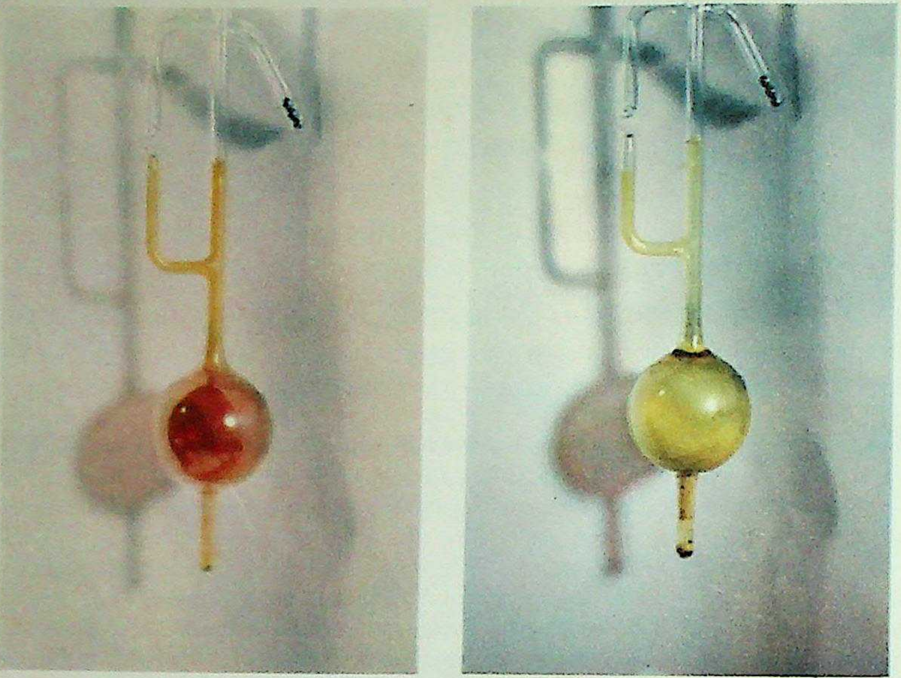
A clear indication that it should be possible to oxidize at least the heavier noble gases to produce salts arose from the discovery of the remarkable oxidizing properties of platinum hexafluoride [11]. N. Bartlett and D. H. Lohmann had established that this gas combined with molecular oxygen in a 1 : 1 molar ratio, forming a red solid. As prepared at low temperatures, this compound proved to be isomorphous with potassium hexafluoroplatinate (v), $K[PtF_6]$. (The symbol (v) refers to the valency of, in this case, the platinum). When formed at a higher temperature, it is isomorphous with potassium hexafluoroantimonate (v), $KSbF_6$. This, together with the magnetic and chemical properties of the oxygen compound, left no doubt that it was dioxygenyl hexafluoroplatinate (v), $O_2^+[PtF_6]^-$. Since the first ionization potential of molecular oxygen (12·2 eV) is almost the same as that of xenon, this element also appeared to be within the oxidizing range of platinum hexafluoride. The lattice energy of a salt is generally inversely proportional to the sum of the ionic radii, but the lattice energy of the hexafluoroplatinates is rather insensitive to cationic radii changes, because the anion is large. A computation made for the hypothetical oxidized xenon compound $Xe^+[PtF_6]^-$ showed that the lattice energy would be little more than 10 kcal less than that of the dioxygenyl salt. The predicted interaction of xenon and platinum hexafluoride was confirmed in a simple and visually dramatic experiment [1]. When deep red platinum hexafluoride vapour was mixed with an excess of xenon gas at room temperature, a yellow solid was immediately precipitated (figure 1). The quantity of xenon consumed was commensurate with a combining ratio of 1 : 1, and accordingly the compound was formulated as xenon hexafluoroplatinate (v), $Xe^+[PtF_6]^-$.

This work aroused much interest at the Argonne National Laboratory, because platinum hexafluoride and the other powerfully oxidizing hexafluorides RuF_6 and RhF_6 were first made there [12]. The Argonne workers confirmed the reaction between xenon and platinum hexafluoride, and then investigated the reaction of xenon with other hexafluorides. They observed that ruthenium hexafluoride reacted with xenon, but that more than two moles of the hexafluoride reacted with each mole of xenon [3]. Further, some reduction of the ruthenium seemed to occur. This suggested that the hexafluoride was, at least in part, acting as a fluorinating agent, and that a fluoride, or some fluorides, of xenon existed. This supposition was very quickly proved correct.

Preparation and chemistry of xenon fluorides

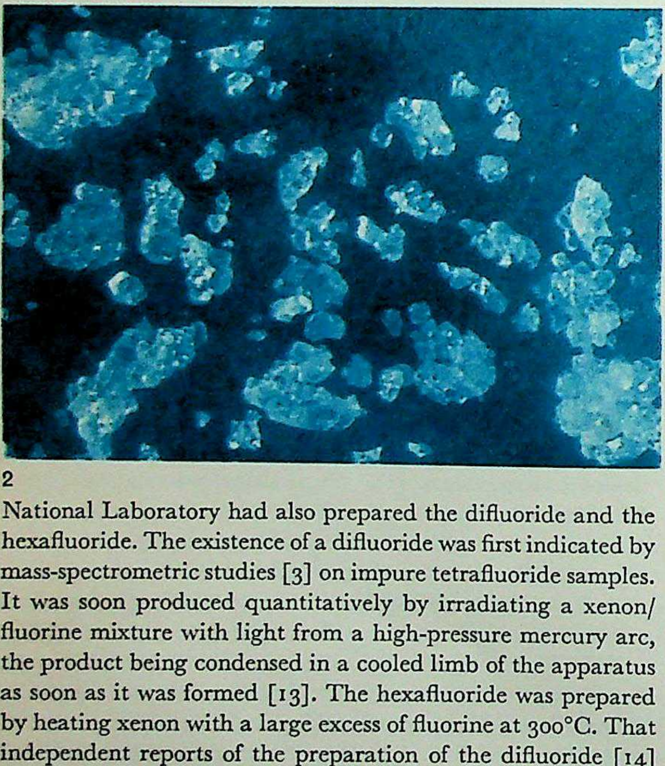
By heating five volumes of fluorine with one volume of xenon to 400°C in a nickel can, followed by rapid cooling in cold water, H. H. Claassen, H. Selig, and J. G. Malm [2] succeeded in making the first fluoride of xenon, xenon tetrafluoride (figure 2). Within a few weeks, they and their collaborators at Argonne

N. Bartlett, *B.Sc., Ph.D., F.R.I.C.*
Was born in 1930 and educated at King's College, Durham. After a year as senior chemistry master at the Duke's School, Alnwick, Northumberland, he became a lecturer in chemistry at the University of British Columbia, where he is now an associate professor. His prime research interests are in inorganic fluorine chemistry, particularly the structural aspects of the subject, and in correlating molecular structure of new compounds with their electronic configuration.



1

Figure 1 The xenon–platinum hexafluoride reaction. Left: Platinum hexafluoride vapour separated from xenon by a glass diaphragm. Right: The product ($XePtF_6$) of the xenon, platinum hexafluoride reaction immediately after the breakage of the glass diaphragm.



2

Figure 2 Crystals of xenon tetrafluoride, seen against a blue background.

National Laboratory had also prepared the difluoride and the hexafluoride. The existence of a difluoride was first indicated by mass-spectrometric studies [3] on impure tetrafluoride samples. It was soon produced quantitatively by irradiating a xenon/fluorine mixture with light from a high-pressure mercury arc, the product being condensed in a cooled limb of the apparatus as soon as it was formed [13]. The hexafluoride was prepared by heating xenon with a large excess of fluorine at 300°C. That independent reports of the preparation of the difluoride [14] and of the hexafluoride [15] appeared simultaneously within three months of the announcement of xenon tetrafluoride testifies to the great interest aroused by that discovery.

At room temperature, all three xenon fluorides are white crystalline solids. The hexafluoride is the most volatile, having a vapour pressure of about 30 mm Hg at 25°C. It melts at 46°C, and the vapour is, surprisingly, pale yellow. The difluoride and tetrafluoride are less volatile, each having a vapour pressure of about 3 mm Hg at 25°C. The xenon fluorides are soluble in liquid hydrogen fluoride, but only the hexafluoride solution is appreciably ionized, presumably as a result of the reaction [16]

$$HF + XeF_6 \rightarrow XeF_5^+ + HF_2^-$$

The tetrafluoride solution fluorinates bright platinum to platinum tetrafluoride. All the fluorides ought to be good fluorinating agents, and it is possible that they may prove to be useful as such, because the reduction product, being xenon, is unlikely to contaminate the fluorinated material. The hexafluoride is thermally less stable than the tetrafluoride, and much more reactive. It appears that the tetrafluoride may be stored indefinitely in dry silica vessels, whereas the hexafluoride reacts rapidly to produce xenon oxide tetrafluoride:

$$2XeF_6 + SiO_2 \rightarrow SiF_4 + 2XeOF_4$$

Compounds of radon and krypton

Efforts were also made at this time to prepare compounds of other noble gases. Radon, in particular, would be expected to form compounds even more readily than xenon. Despite considerable experimental difficulties, P. R. Fields, L. Stein, and M. H. Zirin [17] have established the existence of radon fluoride. Although radon fluoride is apparently much less volatile than any of the xenon fluorides, the very small scale of the preparation does not exclude the possibility of the entire sample being either adsorbed on the nickel reactor surface, or else chemically combined with the protective coating of nickel fluoride. Large-scale preparations, which are technically difficult because of the radiation problem, are in hand and should solve this problem.

A. V. Grosse and his co-workers [18] were able to isolate krypton tetrafluoride by passing an electric discharge through krypton/fluorine mixtures at liquid-nitrogen temperature. The compound is much less stable than its xenon analogue, rapidly decomposing into the elements at 60°C. It is a transparent colourless solid, having a vapour pressure of 115 mm at 25°C.

J. J. Turner and G. C. Pimentel [16] photolysed a solid mixture of fluorine and krypton in an argon matrix at 20°K and obtained a chemical species which infra-red evidence showed to be krypton difluoride. This method, known as the matrix isolation technique, gave no evidence for the formation of an argon fluoride when a mixture of solid argon and fluorine was irradiated alone.

The molecular structures of the fluorides

X-ray and neutron-diffraction studies of the solid xenon difluoride and tetrafluoride have been made [19], and X-ray analysis has revealed a 1 : 1 adduct of the di- and tetra-fluorides [20]. In the case of xenon hexafluoride, neither the X-ray nor the neutron-diffraction studies are complete. Infra-red and

Raman spectra of the two lower fluorides have been unambiguously interpreted [21], but the hexafluoride spectra, although superficially similar to the spectra of molecules with O_h symmetry, exhibit irregularities suggestive of lower molecular symmetry [16]. In both the solid and gaseous state, the difluoride is a linear molecule. In the solid, the molecules lie parallel to one another in a body-centred tetragonal lattice array; the Xe–F distance is 2·00 ± 0·01 Å. Xenon tetrafluoride molecules show square planar coordination of the xenon by the fluorine atoms, in both the solid and gaseous states. The angle between adjacent fluorine ligands is 90°, and the bond length is 1·953 ± 0·002 Å. Infra-red evidence [16] shows that krypton difluoride has a molecular structure similar to its xenon analogue, and it is probable that the tetrafluoride will prove to have the same structure as xenon tetrafluoride.

Oxyfluorides and oxides

Xenon oxide tetrafluoride, which results from the reaction of the hexafluoride with glass, is best made [24] by simply mixing the hexafluoride with an equimolar quantity of water:

$$XeF_6 + H_2O \rightarrow XeOF_4 + 2HF$$

It is a clear, colourless, mobile liquid (m.p. −28°C). The infra-red and Raman spectra have been interpreted [16] on the basis of a model having C_{4v} symmetry, the oxygen atom being at the apex of the square pyramid and the xenon atom being almost coplanar with the four fluorine atoms.

With the fluorides well established, attempts were promptly made to prepare the oxides. Although synthesis from the elements proved to be fruitless, and pointed to the oxides being endothermic, there was an indication that a hydroxide or oxide was produced in the hydrolysis of the tetrafluoride [2]. An oxide was in fact obtained by the hydrolysis of this fluoride, and proved to be the trioxide [22]. When the first samples were obtained there was some uncertainty about the nature of the product. However, it was proved to be the trioxide by X-ray analysis [23], which showed that the material closely resembled iodic acid. Like IO_3^-, the molecule is trigonal pyramidal; the Xe—O bonds are short (1·76 Å) and so must have appreciable double-bond character. The stoichiometry of the reaction is represented approximately by the equation

$$6XeF_4 + 12H_2O \rightarrow 4Xe + 2XeO_3 + 3O_2 + 24HF$$

Although the mechanism is not understood, the xenon trioxide clearly arises by disproportionation of solvated Xe (IV) species. The disproportionation resembles those of the halogen oxy-anions, for example:

$$3ClO_2^- \rightarrow Cl^- + 2ClO_3^-$$

The trioxide is best made by the slow and complete hydrolysis of the hexafluoride [24]:

$$XeF_6 + 3H_2O \rightarrow XeO_3 + 6HF$$

It is non-volatile and dangerously explosive, being comparable in the latter respect to T.N.T. The explosive nature of the compound is compatible with its being highly endothermic ($\Delta H_f{}^\circ{}_{\text{cryst.}} = +96$ kcal mole^{-1}).

Some thermodynamic considerations

The thermal cycle

$$\begin{tikzcd}
N_{(g)} \arrow[d] & + & \frac{1}{2}yX_{2(g)} \arrow[d, "y\Delta H_{f(g)}{}^\circ"] \arrow[r, "\Delta H_{f(g)}{}^\circ"] & NX_{y(g)} \\
N_{(g)} & + & yX_{(g)} \arrow[ur, "y \times \text{average bond energy}"'] &
\end{tikzcd}$$

shows that the formation of a compound from atoms of a noble gas (N) and of a combining element (X), will be exothermic if the average bond energy of the compound exceeds the enthalpy of formation of the gaseous atom X. For thermodynamically stable compounds, under ordinary conditions of temperature and pressure, the average bond energies should therefore exceed, for fluorides, 18·3 kcal; for oxides, 59·2 kcal (double bond); and for chlorides, 29·01 kcal. From the thermodynamic data obtained for the tetrafluoride, the hexafluoride, and the trioxide (Table 1), the average bond energies for the two fluorides are seen to have the same value of 32 kcal, within experimental error. The average energy of the oxygen–xenon bond, which we can presume to be double, is less than this. If the average bond energies in the oxyfluorides are given these values, then xenon oxide tetrafluoride should be an exothermic compound with $\Delta H_f{}^\circ{}_{(g)}$

$$= -\{[(4 \times 32) + 28] - [4 \times 18{\cdot}3(\Delta H_f{}^\circ{}_{(g)}(F)) + 59{\cdot}2(\Delta H_f{}^\circ{}_{(g)}(O))]\}$$

$$= -24 \text{ kcal mole}^{-1}$$

Table 1

Thermodynamic data for XeF_4, XeF_6, and XeO_3

Compound	$\Delta H_f{}^\circ{}_{\text{cryst.}}$	$\Delta H_{\text{sub.}}$	$\Delta H_f{}^\circ{}_{\text{gaseous atoms}}$	Average bond energy
XeF_4	−70[1]; −75[2]	15[3]	128; 133	32–33 kcal
XeF_6	−88[2]	14·9[4]	183	30·5
XeO_3	+95[4]	—	<84	<28

[1] L. Stein and P. Plurien [16].
[2] S. R. Gunn and S. M. Williamson [29].
[3] J. Jortner, E. O. Wilson, and S. A. Rice [16].
[4] C. L. Chernick [16].
[5] S. R. Gunn [16].

However, xenon dioxide difluoride should be endothermic, since

$$\Delta H_f{}^\circ{}_{(g)} = -\{[(2 \times 32) + (2 \times 28)] - [(2 \times 18{\cdot}3) + (2 \times 59{\cdot}2)]\}$$
$$= +35 \text{ kcal mole}^{-1}$$

This oxyfluoride, if it can be prepared, should prove to be a thermally unstable compound.

It is interesting that a plot of average bond energy against atomic number for the series of fluorides SbF_3, TeF_4, IF_5, XeF_6 shows a smooth relationship[1], as represented in figure 3. The values for the highest fluorides of antimony, tellurium, and iodine lie on a similar smooth curve. Extrapolation shows that the average bond energy of an octafluoride molecule ought to be approximately 24 kcal, which would mean a standard heat of formation of the gaseous molecule of −45 kcal mole^{-1}. It is possible that J. Slivnik and his co-workers [16], who have described a yellow product which could be pumped from a bath at −78°C but was retained at −196°C and had a fluorine to xenon ratio of 7·9 : 1, have already obtained this compound. The volatility is, however, greater than one would expect for an octafluoride.

The difficulty of formation and the thermal instability of krypton tetrafluoride are consistent with an average bond energy value of 18–19 kcal. This is obtained by extrapolation from values calculated for the fluorides of arsenic, selenium, and

---

[1] The decrease in average bond energy, with increasing coordination number, does not mean that the intrinsic bond energy is decreasing; in fact, the reverse may be true. For the derivation of intrinsic bond energies, the valence-state promotional energies (which would be expected to increase with coordination number) for the reacting gaseous species would need to be known.

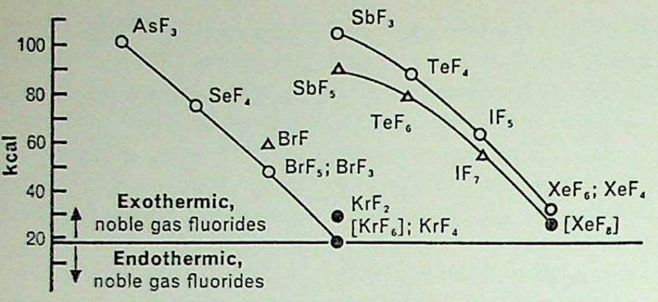


Figure 3 Average bond energy (kcal) for some fluorides. Square brackets indicate unknown compounds, open circles, experimental values; closed circles, hypothetical values.

bromine (figure 3). In view of the greater bond energy of bromine monofluoride (59·4 kcal) compared with the trifluoride and pentafluoride (48·3 and 47.7 kcal respectively), it is possible that the average bond energy for krypton difluoride may be appreciably greater than 19 kcal, making this a thermally much more stable compound than the tetrafluoride. It is of interest that Turner and Pimentel [16], by their matrix isolation technique, obtained evidence for the existence of only the difluoride.

For oxides and chlorides of the noble gases to be formed exothermally, the average single-bond energy needs to be greater than ~ 30 kcal because $\Delta H_{f(g)}{}^{\circ}$ (O) $= 59{\cdot}2$ kcal/mole and $\Delta H_{f(g)}{}^{\circ}$ (Cl) $= 29{\cdot}01$ kcal/mole. Consideration of the average bond energies, estimated from the available thermodynamic data for the oxides and chlorides of antimony, tellurium, and iodine (figure 4), shows, however, that the bond energies for xenon oxides and chlorides should be less than this. Consequently, the xenon oxides and chlorides should be endothermic compounds. This has been confirmed for the trioxide, which exists because of kinetic stability. It is therefore not surprising that attempts to make chlorides by synthesis from the gases failed, and an attempt [25] to prepare the chloride by interaction of boron trichloride with xenon tetrafluoride at −78° also failed:

$$3XeF_4 + 4BCl_3 \rightarrow 3Xe_{(g)} + 6Cl_2 + 4BF_{3(g)}$$

The chlorides may not be kinetically stable, and hence it may not be possible to prepare them under ordinary conditions of temperature and pressure.

The nature of the bonding and of the molecular shape

There is no doubt that the chemical bonding in the noble gas compounds is similar to that found in other compounds. In particular, the bonding must be similar to that found in the inter-halogen compounds, the halogen oxides, and other halogen compounds. For example, the bond lengths, bond strengths, and molecular shapes of the fluorides of xenon are comparable to those of iodine fluorides. Since the molecular shape of xenon tetrafluoride is the same as that of the iso- or pseudo-isoelectronic ions $[ClF_4]^-$, $[BrF_4]^-$, and $[ICl_4]^-$ and the shape of the difluoride the same as $[ICl_2]^-$, it is to be expected that XeF_6 will be similar to the $[IF_6]^-$ ion. Unfortunately, the structure of the latter is unknown. There is, however, evidence that the non-bonding electron pair is sterically active, because the hexafluoroiodates (v) are not isomorphous with those hexafluorometallates that do not contain non-bonding valence electron pairs, for example $KSbF_6$. Simple valence electron-pair repulsion theory predicts molecular shapes for the noble gas compounds that are consistent with the experimental findings (figure 5) [16]. Thus the difluoride is predicted to be a linear molecule having three non-bonding electron pairs at the points of an equilateral triangle, situated in the equatorial plane of the molecule. The tetrafluoride is predicted to be square planar, having one non-bonding electron pair above and one below the plane, both on the fourfold axis of the molecule. The hexafluoride should be a slightly distorted octahedron, having a non-bonding electron pair at the centre of a face or at the mid-point of an edge. In the xenon oxide tetrafluoride molecule, the doubly bonded oxygen is predicted to lie on one side of the plane containing the xenon atom (which is square coordinated by four fluorine atoms) and the non-bonding electron pair on the other. The fluorine atoms and the xenon are predicted to be almost coplanar, because of the near equality in repulsive power of a double bond and an electron pair. The trigonal pyramid shape of the XeO_3 molecule is seen to be consistent with the presence of a non-bonding electron pair at the apex of the pyramid, this pair repelling the bonding electrons (four in each Xe═O bond) to produce the pyramid.

In this treatment, which is not capable of being made quantitative, all the valence electrons are treated equally, and it is implicit in the theory that *d*-orbitals are involved in the bonding. R. E. Rundle [26] and J. Hinze and K. S. Pitzer [16], however, have argued that *d*-orbital involvement should be small. Indeed, the most popular theoretical model has been the simple molecular orbital scheme; this uses *p*-atomic orbitals and was first used by Pimentel [27] to discuss inert gas

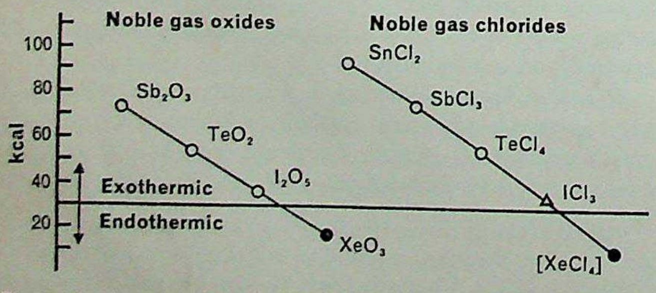


Figure 4 Average bond energies (kcal) for some oxides and chlorides. Square brackets indicate unknown compounds; open circles, experimental values; closed circles, theoretical values.

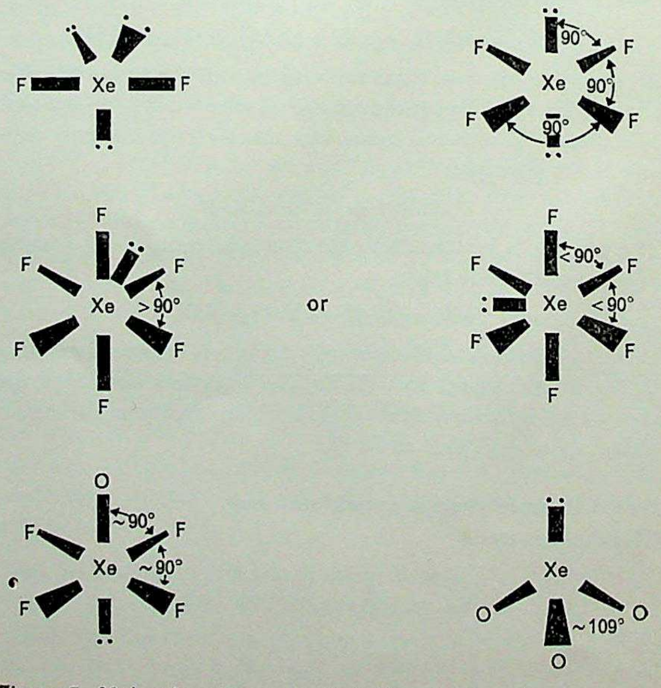


Figure 5 Molecular shapes predicted by simple valence electron-pair repulsion theory [30].

compounds qualitatively. For the xenon fluorides, Rundle has proposed linear, three-centre, four-electron bonds (F—Xe—F) in which two electrons occupy a bonding molecular orbital, ψ_{μ}, and two occupy a non-bonding molecular orbital, ψ_{δ}. The latter places the electrons mainly on the fluorine atoms, and consequently the bond must be considered to have some ionic character. Since this situation allows only one bonding pair of electrons for the two F—Xe bonds, the bond length is predicted to be greater than for a conventional electron-pair bond. This theory allows for the existence of the three xenon fluorides, XeF_2, XeF_4, and XeF_6, and correctly predicts the molecular shape of the di- and tetrafluorides, Since the linear F—Xe—F bonds should be approximately at right angles, it predicts an octahedral XeF_6.

As has been pointed out by L. C. Allen [16], the controversial theoretical situation should become clearer when the structure of xenon hexafluoride is known. Since there is an accumulating body of evidence to show that this compound is not octahedral, the case for *d*-orbital involvement in the bonding grows stronger. The preparation of an octafluoride would also prove this involvement.

Octavalent xenon

Although the octafluoride is not yet known, the anticipated octavalence of xenon [8] has been attained with the isolation of the perxenates. The reaction of xenon hexafluoride with sodium hydroxide was shown by J. G. Malm and his colleagues [16] to give white crystalline solids which contain xenon in the +8 oxidation state. Several hydrates were stated to exist, but the one that crystallizes first has recently been shown by W. C. Hamilton, J. A. Ibers, and D. R. Mackenzie [28], by X-ray structural analysis, to be $Na_4XeO_6.8H_2O$. The perxenate ion $(XeO_6)^{4-}$ is nearly regularly octahedral, having a mean Xe—O bond length of 1·875 Å. This bond is 0·12 Å longer than the average Xe—O distance in XeO_3 [23], which compares well with the increase of 0·11 Å in the I—O distances in the isoelectronic iodine-containing ions IO_3^- (1·81 Å) and IO_6^{5-} (1·93 Å). Presumably, all the electrons of the xenon outer octet are involved in the $[XeO_6]^{4-}$ bonding.

Conclusion

Approximately one hundred scientific papers have been published since the report of the first xenon compound in 1962. Never before has the picture of the chemistry of an element and the structural properties of its compounds been so rapidly developed. The limitations of the chemistry of the noble gases can now be clearly seen. It is evident that compounds of helium, neon, and argon are unlikely to be produced other than as highly unstable species; even the synthesis of a fluoride of argon appears unlikely. Although more compounds of xenon and radon may be prepared, those of xenon will probably be confined to substituted fluoride molecules (for example, CF_3XeF_5 and $ClXeF_5$); to adducts of the fluorides; to Xe (I) complexes related to $Xe^+[PtF_6]^-$; and to oxygen-containing salts. Radon may well form a chloride and even a more stable oxide than does xenon.

References

[1] Bartlett, N. *Proc. chem. Soc., Lond.*, 218, 1962.
[2] Claassen, H. H., Selig, H., and Malm, J. G. *J. Amer. chem. Soc.*, 84, 3593, 1962.
[3] Chernick, C. L. *et al. Science*, 138, 136, 1962.
[4] Villard, P. *C. R. Acad. Sci., Paris*, 123, 377, 1896.
[5] Travers, M. W. 'Sir William Ramsay', p. 146. Edward Arnold (Publishers) Ltd, London. 1956.
[6] Oddo, G. *Gazz. chim. ital.*, 63, 380, 1933.
[7] von Antropoff, A. *Z. angew. Chem.*, 37, 217, 1924.
[8] Pauling, L. *J. Amer. chem. Soc.*, 55, 1895, 1933.
[9] Ruff, O. and Menzel, W. *Z. anorg. Chem.*, 213, 206, 1933.
[10] Yost, D. M. and Kaye, A. L. *J. Amer. chem. Soc.*, 55, 3890, 1933.
[11] Bartlett, N. and Lohmann, D. H. *Proc. chem. Soc., Lond.*, 115, 1962.
[12] Weinstock, B. *Rec. chem. Progr.*, 23, 23, 1962.
[13] Weeks, I. L., Chernick, C. L., and Matheson, M. S. *J. Amer. chem. Soc.*, 84, 4612, 1962.
[14] Smith, D. F. *J. chem. Phys.*, 38, 270, 1963.
[15] Dudley, F. B., Gard, G., and Cady, G. H. *Inorg. Chem.*, 2, 228, 1963.
Weaver, E. E., Weinstock, B., and Knop, C. P. *J. Amer. chem. Soc.*, 85, 111, 1963.
Malm, J. G., Sheft, I., and Chernick, C. L. *Ibid.*, 85, 110, 1963.
Slivnik, J., *et al. Croat. Chem. Acta*, 34, 253, 1962.
[16] 'Noble Gas Compounds'. H. H. Hyman (editor), University of Chicago Press, Chicago. 1963.
[17] Fields, P. R., Stein, L., and Zirin, M. *J. Amer. chem. Soc.*, 84, 4164, 1962.
[18] Grosse, A. V., Kirshenbaum, A. D., Streng, A. G., and Streng, L. V. *Science*, 139, 1048, 1963.
[19] Levy, H. A. and Agron, P. A. *J. Amer. chem. Soc.*, 85, 241, 1963.
Agron, P. A. *et al. Science*, 139, 842, 1963.
Burns, J. H., Agron, P. A., and Levy, H. A. *Ibid.*, 1209, 1963.
Templeton, D. H., Zalkin, A., Forrester, J. D., and Williamson, S. M. *J. Amer. chem. Soc.*, 85, 242, 1963.
[20] Burns, J. H. Conference on Noble Gas Compounds, Argonne National Laboratory. April 1963.
[21] Agron, P. A., Begun, G. M., Levy, H. A., Mason, A. A., Jones, C. G., and Smith, D. F. *Science*, 139, 842, 1963.
Claassen, H. H., Chernick, D. L., and Malm, J. G. *J. Amer. chem. Soc.*,
[22] Williamson, S. M. and Koch, C. W. *Science*, 139, 1046, 1963.
[23] Templeton, D. H., Zalkin, A., Forrester, J. D., and Williamson, S. M. *J. Amer. chem. Soc.*, 85, 816, 1963.
[24] Smith, D. F. *Ibid.*, 85, 816, 1963.
[25] Bartlett, N. and Jha, N. K. Unpublished work.
[26] Rundle, R. E. *J. Amer. chem. Soc.*, 85, 112, 1963.
[27] Pimentel, G. C. *J. chem. Phys.*, 19, 446, 1951.
[28] Hamilton, W. C., Ibers, J. A., and Mackenzie, D. R. *Science*, 141, 532, 1963.
[29] Gunn, S. R. and Williamson, S. M. *Ibid.*, 140, 177, 1963.
[30] Gillespie, R. J. and Nyholm, R. S. *Quart. Rev. chem. Soc.*, 11, 339, 1957.

Whiskers

J. E. Gordon

The growth of whiskers—thin hair-like crystals—on metals has been known for some time to be a cause of failure in electrical condensers. The whiskers were, originally, only a relatively rare curiosity, but it has been found that they can be prepared from an extremely wide range of materials. They are very much stronger than larger crystals. This article discusses the reasons for this strength and the light that it throws on the process of crystallization.

The great majority of solid elements and compounds are crystalline, and the commonly occurring shapes of these crystals are generally compact enough to be described as 'lumps', whatever their actual crystallographic form. Some substances, such as iodine or potassium permanganate, form needle-shaped crystals naturally, but, even so, the ratio of length to breadth of such crystals is usually quite low, perhaps about ten to one. With a very few exceptions, the innumerable crystals that occur in nature and in the laboratory are mechanically very weak. They possess strengths that are far less, perhaps a thousandfold less, than the strengths predicted from the character of their interatomic bonds. There are, on the other hand, thin hair-like crystals that have great mechanical strength, sometimes approaching the theoretical value. These are known as 'whiskers'. 'Metal whiskers' are often mentioned, but in fact whiskers are neither more common nor more interesting in metals than in non-metals. The interest of whiskers lies as much as anything in the universality of the phenomenon. Whiskers do not seem to be confined to any chemical group of elements or compounds, nor to any particular crystallographic system. Furthermore, almost any crystallization process, chemical or physical, from vapour or from solution, can usually be made to yield whiskers by adjusting the conditions of growth.

Traditionally, high strength existed only in a mere handful of materials—perhaps half a dozen metals and a few biological substances such as collagen and cellulose. Recently, glass fibres and certain plastics have been added to these, but on the whole, the structural engineer is still very limited, especially in the choice of materials with high Young's modulus and high temperature-resistance. It now seems to be possible that, either by exploiting whiskers directly or, more probably, by using the increased knowledge that comes partly from the study of whiskers, of the relationship between crystal growth and mechanical strength, we may be able to get high strength from any crystalline solid, irrespective of its chemical nature. The engineering potentialities of this kind of knowledge would of course be very important.

Whiskers range in thickness from a few hundred Ångström units upwards, but typically they are perhaps a micron or two thick, and vary in length from a few microns up to several centimetres. One of the attractive things about whiskers is that much of their growth can be studied, and their strengths can be measured approximately, with the simplest of apparatus. A good deal of the work in my own laboratory has been done with simple optical microscopes and could equally have been done by any amateur. In fact, it is curious that the field has not attracted the spare-time scientist.

The strong presumption is that whisker growth, like that of most crystals, is usually dependent on screw dislocations (figure 1), especially since one or more lines, that perhaps correspond to the dislocation cores, can sometimes be seen running down the middle of transparent whiskers. It is well known that crystals do not grow by the alighting of molecules in a random manner on the surface, like swallows on a telephone wire. Rather, they grow by the regular advance of a step resembling that which exists in a course of brickwork when a wall is being built. It can be shown that it is energetically much more favourable for new molecules to come to rest in the notch afforded by the step than to assume an isolated position. New growth thus spreads systematically as a layer across the old surfaces. In the ordinary way, such a process would clearly need nucleating afresh for each molecular layer, because it would come to an end when a layer reached the edges of the crystal surface. F. C. Frank therefore suggested [1] that the growth layer is really a spiral, and that, once nucleated, the growth layer runs round and round. In normal big crystals there are large numbers of screw dislocations that may interact in a complicated manner. In a whisker there is probably one or, at most, perhaps a dozen. The majority of whiskers can thus be thought of as a rather flat spiral of atoms or molecules: in atomic terms, they are, of course, of immense length. Most whiskers probably grow by screw dislocations, but the process may occur in either of two main ways, tip growth and root growth: that is to say, the filament may lengthen by adding new material either to the tip or at the root. Tip growth occurs in general when crystallization takes place from solution or from the vapour phase: root growth is mainly a solid-state phenomenon, and is fairly rare.

Root growth

Since relatively little is known about it, root growth may be dismissed fairly briefly. When a ductile material, such as a metal, is subjected to a compressive mechanical stress, whiskers will sometimes grow out of it (figure 2). This may occur, for instance, if lead or tin is clamped in a vice for a few weeks. It seems that atoms diffuse from the interior through the solid material to a screw dislocation, and that this metal is then extruded to the extent of perhaps a centimetre or more. The rate of growth is proportional to the applied stress, but is in any case quite slow, seldom as much as a millimetre a week. Since the energy to carry out the growth process comes from the strain energy of the surrounding material, surfaces with locked-in stresses, such as may be induced by a thin coating of another metal, are susceptible to this type of growth. This is why plated electrical parts, such as contact breakers, are sometimes short-circuited by the growth of whiskers. Conductors in electric cables also occasionally grow whiskers that force their way through thick plastic insulation. There does not seem

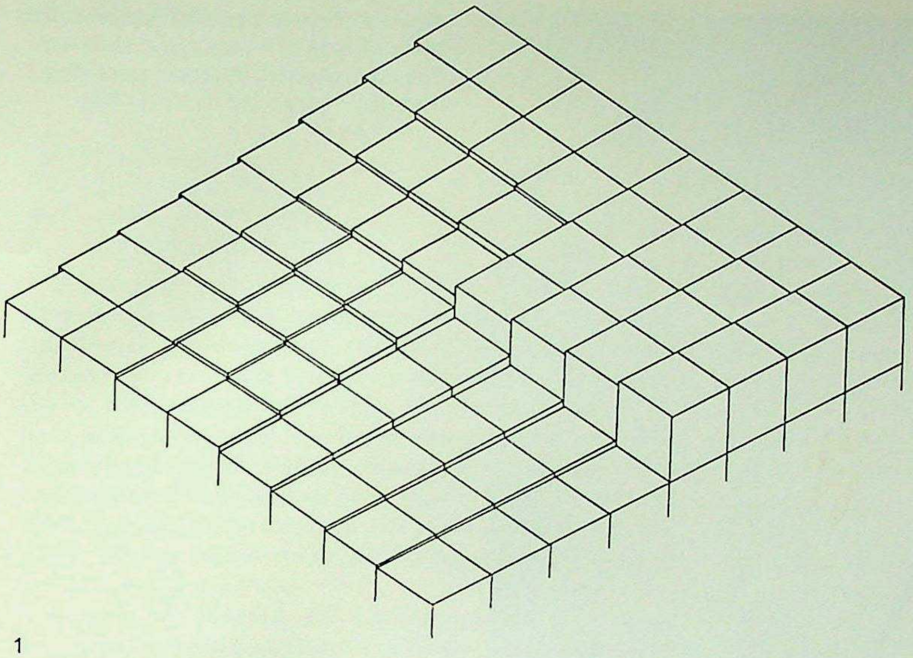
J. E. Gordon, *B.Sc.*

Was born in 1913 and educated at the University of Glasgow. From 1938 to 1955 he was at the Royal Aircraft Establishment, Farnborough, where he became head of the Plastic Structures Group; and from 1955 to 1962 he led a research group at Tube Investments Research Laboratories, Hinxton Hall, working on the strength of non-metallic materials. This work is being continued for the Ministry of Aviation at the Explosives Research and Development Establishment at Waltham Abbey, where he is superintendent of a group that is also concerned with high-temperature application of two-phase systems.

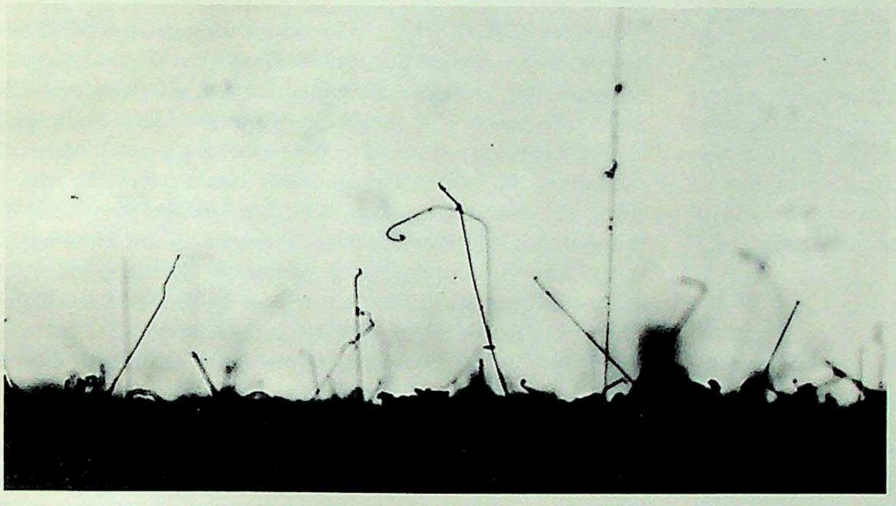
Figure 1 Screw dislocation [1].

Figure 2 Tin whiskers produced by root growths. (×50)

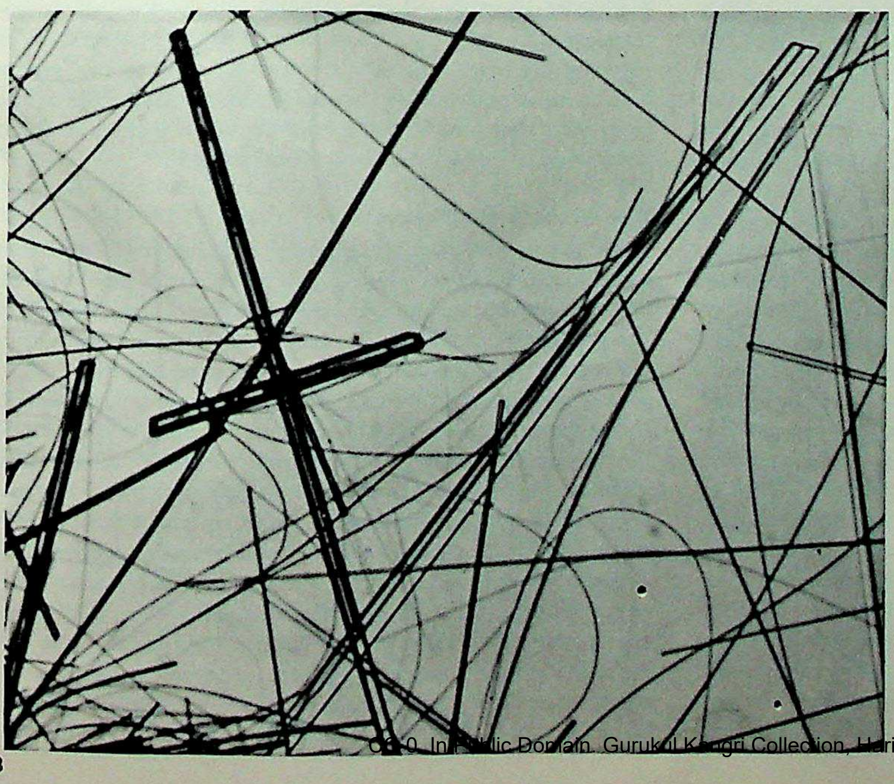
Figure 3 Hydroquinone whiskers growing from aqueous solution on a microscope slide. (×50)



1



2



3

to be any effective practical insurance against this trouble. The process of root growth is so slow and uncertain that it is difficult to study, and it is probably useless as a practical means of whisker production.

Tip growth is much faster, much commoner, and, since it is controllable, probably more useful. The phenomenon can be easily studied by crystallizing a suitable material, such as hydroquinone, from aqueous solution under a bench microscope at room temperature (figure 3). The mechanisms observed in this case are generally similar to those that occur when other substances crystallize from solution, and also during crystallization from vapour.

Growth from solution

When the solvent is allowed to evaporate or to cool slowly, the first things normally visible are fine filaments. These spread rapidly through the liquid, with a rate of growth sometimes as high as a millimetre per second and a thickness in the region of a micron. The filaments are often bent and tangled like cotton wool. Observation shows that the filaments are naturally straight if they are allowed to take up their preferred relaxed position. They are, however, so thin that they are very flexible, and they are no doubt blown this way and that by Brownian movement and other disturbances as they grow. These first filaments we have called 'leaders', and if they are observed in the electron microscope, they are generally found to be extremely smooth down to something approaching atomic dimensions. In the next stage of growth, the leaders thicken by the spreading down them of growth layers, which are often thick enough to be seen clearly in the optical microscope. It is not certain how these layers are nucleated, but C. C. Evans has suggested that this occurs when the leaders touch each other, which they must do from time to time. Sometimes the growth layers encircle the whole leader like a sleeve, while sometimes they seem to occur separately

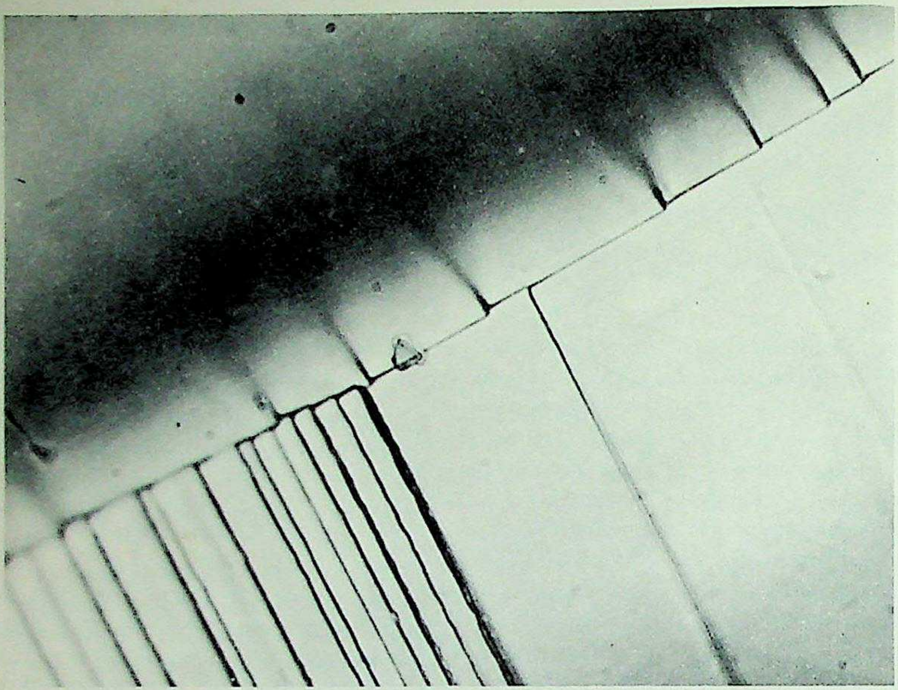


Figure 4 Growth layers on adjacent faces of large antimony oxide whisker. The edge at which the faces meet runs diagonally across the figure. (×550)

on each face (figure 4); but in any case the crystal thickens, and the final product is a needle crystal of the familiar kind, which presents no special features of interest. The interesting stages are the early ones, and it is these that yield the true whiskers. These early stages of growth have a great many variants and are most beautiful and instructive to watch, especially by anyone who is interested in elasticity, because of the light that is thrown upon the mechanisms of strength and fracture.

Tensile-strength measurement

There are several experimental ways of ascertaining the tensile strength of a material. The simplest in theory, and the most convincing, is to pull a rod-like specimen asunder in a tensile-testing machine. This is much more difficult to do reliably in practice than it sounds, and it becomes very difficult indeed with specimens of the diameter of whiskers. An acceptable alternative, at least for fully elastic materials, is a bend test. If a rod of thickness t is bent to a radius r, it follows from the simplest of geometry that there is a strain, $\varepsilon = t/2r$, at the surfaces of the rod. Hence if the rod is made of a material with a known Young's modulus, E, then the tensile stress at the surface is $\sigma = \varepsilon E$. A bend test is unsatisfactory when the material is not fully elastic—that is to say, if it flows plastically before failure, as ductile materials do—or if the specimen is so very thin that it is not easy to measure the thickness accurately.

It is not too difficult to do bend tests on whiskers under the microscope so as to give results of fair accuracy. There are two methods. In the simplest method, the whisker is mounted, preferably in grease or jelly, on a microscope slide, and it is then bent with a dissecting needle while under observation. Bend radii can be checked photographically or by projection on a ground glass screen. After failure, the two broken halves should spring straight if no plastic flow has occurred—that is, if the deformation was fully elastic. A less clumsy method is to manipulate the whisker into a bent form while it is still growing, and then to allow it to thicken at a constant bend radius until it breaks. There are limitations to both methods. One of these is the difficulty of measuring, in an optical microscope, the thickness of filaments whose dimensions are comparable with the wavelength of light. For this reason the method becomes very inaccurate for the thinner whiskers, of diameter less than about 2 μ. This is one of the reasons why there have been many attempts to devise machines for breaking whiskers in direct tension, and these attempts have had varying degrees of success. In practice, the difficulty is not so much in attaining the required degree of sensitivity in measuring small loads and extensions, as in getting consistent results. In this respect electrical devices are often defective because they tend to drift—that is, they need frequent correction—and are sensitive to the level of electrical noise. Probably the most successful micro-testing machine is that divised by D. M. Marsh [2]. The operation of this machine is wholly mechanical and optical, and it is remarkably free from drift. The Marsh machine will measure extensions repeatably to about 5 Å and will give satisfactory load/extension curves for specimens as small as one tenth of a square micron in cross-section and about half a millimetre in length. Such machines open up a new field of research in investigating the strength of very small bodies. Incidentally, almost the only feature that all micro-tensile testing machines have in common is the method of gripping the specimen; this is 'soldered', often to two quartz lugs, with a trace of diphenylcarbazide, a most convenient compound that wets almost everything, melts sharply at about 180°C and, when hard, has excellent strength.

The strength of whiskers

We may now describe the results of the strength tests that have been made upon whiskers and consider the way in which these results throw light upon the problems of the mechanical strength and weakness of solids of all kinds. Of the various ways of estimating the theoretical strength of materials, the simplest is that put forward by A. A. Griffith over 40 years ago [3]. He suggested that since, when a material fails, two new surfaces must be created, the free surface-energies of the new surfaces must be provided by the strain energy of the two layers of atoms that have been pulled apart. A minor difficulty is that we cannot calculate this strain energy very accurately since we know the shape of the interatomic force curve at large strains only approximately. Although the surface energy of a solid is seldom known very precisely, it can be estimated by extrapolating from measured values for the same material in the liquid state. However, the latitude for error from both these approximations is not large enough to affect the dependent arguments in any important way. One way of writing down the Griffith hypothesis is that due to M. Polanyi and E. Orowan, who stated [4], that

$$\sigma = \sqrt{(2\alpha E/a)}$$

where σ = breaking stress

E = Young's Modulus

α = free surface energy

a = interatomic distance

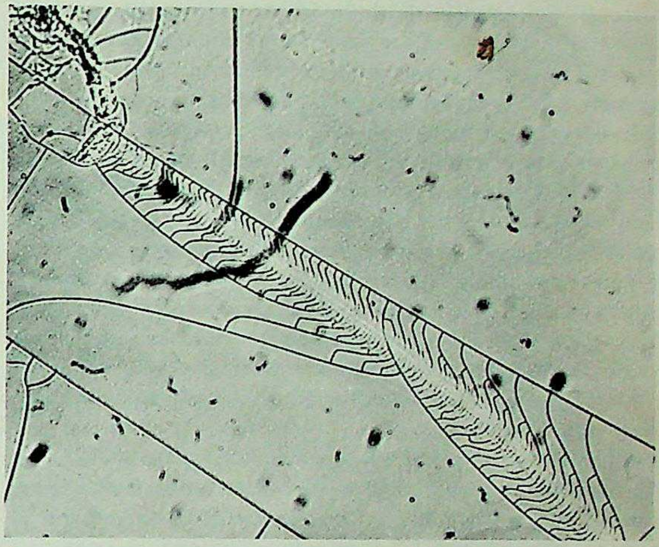
Whatever method of calculation is used, we generally get a figure for the theoretical strength in the region of 0·20E and, what is even more interesting, this figure remains very roughly the same for a wide range of elements and compounds, irrespective of their chemical composition. In other words, one ought to be able to stretch the interatomic bonds in most materials, elastically and without fracture, by something like 20 per cent, and therefore stretch the material as a whole to the same extent. Even if generous allowances for the various uncertainties are made, this figure must at any rate lie somewhere between 10 per cent and 30 per cent.

Actually, the observed elastic breaking strains of the great majority of crystalline and glassy solids lie in the range 0·01–0·1 per cent of E—between a thousandth and a hundredth of the theoretical value. A very high tensile steel may have a value of slightly over 1·0 per cent of the Young's modulus. In his classical paper [3], Griffith showed that, whereas ordinary bulk glass has strengths that vary between say 0·01 and 0·05 per cent of the modulus, the same glass drawn into thin fibres may show strengths of between about 5 and 8 per cent of the modulus. Griffith deduced that the theoretical strength of his glass was in the region of 10 per cent of the modulus. We should today consider this value to be on the low side but the error, if there is one, is not serious. This was probably the first time that the theoretical strength of a material had been both plausibly calculated and shown to be approachable experimentally, and Griffith's paper thus marked an important scientific advance. He showed that, experimentally, thick fibres were nearly as weak as bulk glass and that strength increases as the fibres became thinner in such a fashion that the strength/thickness curve roughly resembled a rectangular hyperbola—that is, the strengths increased very rapidly when the fibres became a few microns thick.

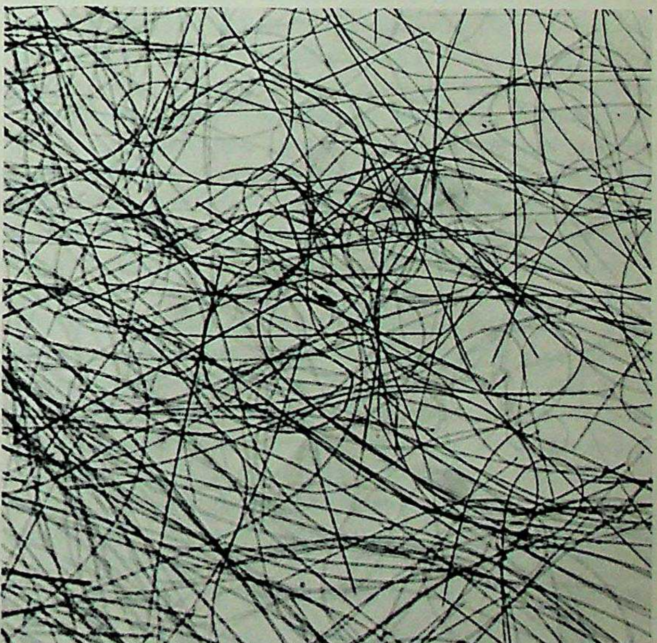
The action of dislocations

Griffith pointed out that the problem was not so much to explain why the thin fibres were strong, for a single chain of atoms must have the theoretical strength, as to explain why thick fibres and bulk glass were weak. To do this, he postulated the existence of very small cracks in the larger specimens. It can be shown that such cracks may introduce very high concentrations of stress and thus precipitate failure. For instance, a crack a micron long, which would be invisible by normal viewing techniques, might reduce the strength perhaps a hundredfold. Griffith seems to have considered that the cracks existed throughout the body of the material and that there was simply no room for them in the thinnest fibres. It was later shown by E. N. da C. Andrade [5], and many others, that crack-like defects normally exist on the surface of glass (figure 5) and that, if the surface is dissolved away, with suitable precautions, glass rods and fibres of any size will exhibit high strengths. There is now an overwhelming amount of evidence that the interior of all glasses exhibits something like the full theoretical strength, but that the practical strength is governed by the presence of very small geometrical defects in the surface, from which cracks penetrate catastrophically to the interior. The failure of normal glasses to reach their theoretical strength is therefore now quite well understood.

The corresponding weaknesses of large ductile crystals has been very adequately and elegantly explained by the dislocation mechanism. As satisfaction at this success was perhaps at its height about the time when C. Herring and J. K. Galt showed [6] that fine whiskers could exhibit very high strengths, these strengths were immediately attributed to the absence of dislocations. Whiskers, it was supposed, were very perfect crystals (figure 6). This explanation, as it turned out, was not completely true, as some whiskers do in fact contain dislocations. However, it was found that whiskers of brittle materials, in which dislocations are in any case immobile, were strong. A mobile dislocation, and still more a dislocation source, of course provides a perfectly plausible weakening mechanism. But an immobile dislocation merely reduces the degree of order in the



5



6

Figure 5 An accidental scrape and natural cracks in the surface of a glass rod (~ ×170). Photograph by Margaret Parratt (*Proc. roy. Soc., A.*, **249**, 65, 1958).

Figure 6 Silicon nitride (Si_3N_4) whiskers (×350). Note smooth surfaces and absence of steps. (Crown copyright.)

crystal, and since glass, which presumably exhibits a maximum degree of molecular disorder, is shown to be very strong, it seems unlikely that the presence of a few immobile dislocations will have a serious weakening effect.

Experimental results

We may now turn to the experimental results on whiskers. Using various types of micro-testing machines, a number of workers have obtained size/strength curves that are roughly hyperbolic in shape and show a marked increase in strength when the thickness of the whisker is reduced to 2–3 μ. The actual strengths vary considerably with the substance chosen, but, plotted as elastic breaking strain or as a percentage of Young's modulus, all the curves show a remarkable similarity. This is true, with minor reservations, for metals and non-metals, and for brittle and ductile materials, and it is not much affected by the way in which the whiskers are grown. Since the high-strength region of the size/strength curve is so steep, it is perhaps too much to expect all the maximum strengths to be consistent, but, very roughly, such whiskers as can be handled in micro-testing machines show maximum strengths of between 3 per cent and 6 per cent of the modulus. The experiments can be extended to thinner fibres by observing the bending of very thin whiskers under the electron microscope. We find a highest published value of 12 per cent strain for forsterite (Mg_2SiO_4) [7] and some rather higher unpublished values. This compares with a confirmed maximum of about 14 per cent for silica glass. Thus the highest experimental strength values, or rather strain values, seem to be approaching the theoretical value of 20 per cent of the modulus. This value, it will be remembered, is an approximate one only.

Since the majority of crystals are brittle, the majority of whiskers are brittle also, and might be expected to behave very much like glass fibres. It is not unreasonable, therefore, to presume the existence of a crack mechanism very similar to that in glass. In fact, however, there is little or no evidence for the existence of Griffith cracks in whiskers. If they do exist, they almost certainly play quite a minor role. What do exist, and can be observed, are growth steps. These steps have already been referred to as part of the mechanism by which thin whiskers become thicker. As a rule, such steps have sharp, 'crystallographic' faces, and in the electron microscope they can usually be seen to have root radii approaching molecular dimensions: one set of measurements gives 40 Å as an average figure. Now the sharp-ended crack is well known to be a dangerous weakening defect; but very little was known about the corresponding step, which might be regarded, as it contains an angle of 90°, as half a crack. Although a few *ad hoc* solutions were available, there was until recently no generalized solution for the stress concentration at a step, although a step at first sight seems much less weakening than a crack. However, D. M. Marsh obtained a general solution [8] for the crack problem and found that, other things being equal, a 90° step was about 85 per cent as weakening as the corresponding crack. This was a rather unexpected result.

Marsh was able to show that there is an excellent correlation between the observed geometry of the steps on whiskers and their measured strength. This appears to afford an entirely adequate explanation of the strength of the majority of whiskers. It remains only to explain why thick whiskers should have higher steps than thin ones. This problem involves an elegant application of M. J. Lighthill's traffic theory. Lighthill [9] was concerned with the clotting of traffic on busy roads and with the manner in which the faster cars accumulate behind the slower ones. It is clear that growth layers, although they can all be supposed to be initially monomolecular, are unlikely all to advance at the same speeds. The actual speeds will be scattered statistically about the mean speed. Since the faster growth layers have no means of passing the slower ones, and because the higher the step, the slower it is likely to advance, as more time is required for the material to accumulate, it is clear that, on average, steps will tend to increase in height with time. The faster layers will simply accumulate on the advancing edge of the slower. Since the thicker crystals will necessarily be older, they will on the average possess higher steps. In fact this mechanism seems very adequately to account for the strength of thin whiskers and the weakness of large ones. A confirmation is given by the experiment of W. C. Dash [10]. He took a large (2 cm) silicon crystal and very carefully polished it. He attained in this way strengths of about 300 tons/in², a value comparable to that found in whiskers.

Thus we see that the reason for the strength and weakness of crystals, and especially brittle crystals, is generally similar in principle to that for glasses. That is to say, the interior is nearly always very strong, perhaps up to the theoretical strength, and failure is precipitated by the existence of small geometrical defects in the surface. However, whereas in glasses the typical surface defect is the Griffith crack, it seems that in crystals the typical defect may be the surface step. In brittle crystals, such steps may initiate cracks that spread catastrophically. In ductile crystals, steps may apparently act as dislocation sources that in turn may set in train a series of dislocation mechanisms leading eventually to mechanical failure.

References

[1] Frank, F. C. *Disc. Faraday Soc.*, No. 5, 48, 1949.
[2] Marsh, D. M. *J. sci. Instrum.*, 38, 229, 1961.
[3] Griffith, A. A. *Phil. Trans.*, *A*, **221**, 163, 1920.
[4] Orowan, E. *Rep. Progr. Phys.*, **12**, 192, 1949.
[5] Andrade, E. N. da C. and Tsien, L. C. *Proc. roy. Soc.*, *A.*, **159**, 346, 1937.
[6] Herring, C. and Galt, J. K. *Phys. Rev.*, **85**, 1060, 1952.
[7] Sears, G. W. *J. chem. Phys.*, **36**, 862, 1962.
[8] Marsh D. M. Proceedings of Seattle Conference on Fracture. (To be published.)
[9] Lighthill, M. J. *Proc. roy. Soc.*, *A.*, **229**, 317, 1955.
[10] Dash, W. C. 'Growth and Perfection of Crystals', p. 189. John Wiley & Sons, London and New York. 1958.

The communication of information

Colin Cherry

The concept of the measurability of information has been clarified in the last decade, and the science of communication theory is now well developed. This article describes the basis of the theory; shows its applicability to the communication of information with the minimum of error; and demonstrates that error cannot be completely eradicated. The theory developed in connection with telegraphy, and some ideas are as old as the Morse Code. It now has wide applications in many other fields.

The gigantic scales of national and world communication networks, and the quantities of messages and data they convey, are indicators of the growth of our world society today. These great networks are derivatives of the Persian and Roman postal services that held those empires together [1]. The statistics of world communication today are fully documented [2] and the use of most services is growing at an increasing pace. Thus the postal services in advanced countries are saturated, and the use of Telex is very rapidly expanding. When we consider the immense capital investment in these networks and their social and economic importance, it seems surprising that the 'commodity' that these networks 'sell' has not been defined and measured until modern times. This commodity is information.

The mathematical theory that has been evolved around the concept of information has implications in many fields of science other than telecommunication, as may seem reasonable, for such concepts as 'communication' and 'information' are very broad indeed. The theory really forms part of epistemology; it has a close relationship too with statistical thermodynamics. But, though a similar theory might have been originated in several other sciences, in fact it was not. For this reason, I shall describe in this article the main aspects of communication theory in relation to telecommunication.

Messages, signals, and information

Broadly speaking, we can be said to gain 'information' in two distinct ways: (*a*) by communication, in some agreed language or sign-system in which messages may be constructed, and (*b*) by observation, as when looking down a microscope. In the latter case, we cannot be said to be 'communicating with nature'; it does not talk to us in language and symbols. Nevertheless, the theories of communication and of observation have much in common [7]. It should however be noted that Sir Ronald Fisher's definition [10] of an 'amount' of statistical information stands quite apart from the theory we are discussing here.

For true communication, the sender and receiver use an agreed set of signs, an alphabet, that has rules whereby the signs can be sequentially ordered into messages. The signs and rules constitute a syntax. The sender successively selects signs and so constructs messages; then he sends physical signals to the receiver which give the receiver the power to select the same signs from his alphabet. This 'selective potential' of the signals constitutes their information content.

Figure 1 illustrates a simple model of a channel of communication. The sender and receiver possess the same alphabet of signs. For convenience we may think of the letters of the English alphabet, or a list of agreed words; the nature of this alphabet must be specified. The sender selects the signs and sends signals that specify these selections. This specification of selections by signals is called a code. Note that the mathematical theory is expressed, not in the language of the sender or receiver, but in a meta-language of an outside observer who sees both.[1]

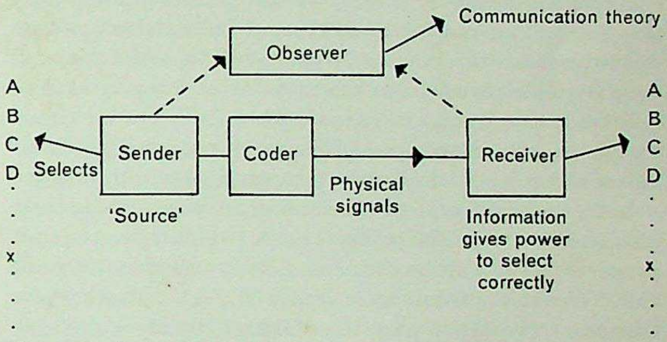


Figure 1 A simple model of a communication channel.

Let us look at this selection and coding process in its simplest possible form. Figure 2 shows, on the left, an alphabet of eight signs represented by the letters A to H of the English alphabet. Let us assume further, for the moment, that the sender selects all these equally often and independently; that is, all $p(x)$ the probability of $x) = \frac{1}{8}$ and there are no 'spelling' rules. Before the receiver receives a signal his doubt is spread uniformly over all the signs. The simplest possible action for the sender is to divide the alphabet into two equal groups (column 1) and to signal one binary-digit to the receiver. Labelling these alternative groups 1 and 0, representing any binary choice, such as yes/no, up/down, dot/dash, the receipt of a 1 instructs the receiver that the next selected sign lies in the upper group. The receiver's doubt, in other words, is halved by the receipt of one binary-digit.

Using such binary division, it requires three digits to specify uniquely each selection out of eight alternatives, as $\log_2 8 = 3$. For example, the signal 101 specifies the selection of C uniquely. Binary signals are of course used in practice in the form, for example, of punched tape. A typical sequence from such a coded source could be . . . 000011111100100111 . . . requiring a single synchronizing point, or origin, since *a priori* all code groups are known to have three digits.

In general, when such a source has N signs, selected at the rate K per second, the rate of transmission of information, H, is

$$H = K \log_2 N \text{ binary-digits/second} \qquad (1)$$

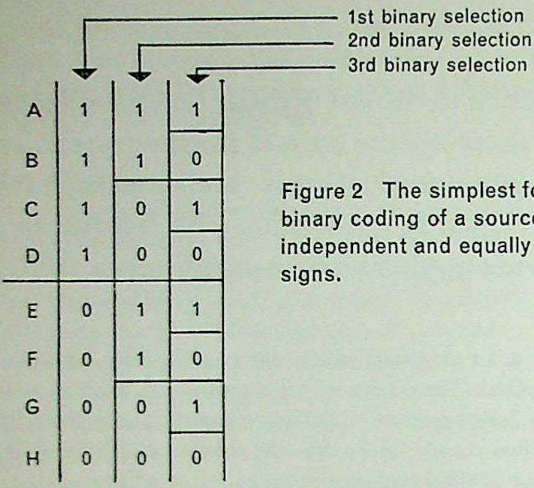
Usually 'binary-digits' is abbreviated to 'bits'. Note, however, that there is nothing magic about the base 2; it is merely the first plural number and a base that has been widely accepted.

Colin Cherry, *D.Sc. (Eng.), A.M.I.E.E., Hon. A.C.G.I.*

Was born in St. Albans in 1914 and educated at St. Albans School and at Northampton Polytechnic, London, while working as laboratory assistant at the Research Laboratories of the General Electric Co. Ltd, Wembley. He subsequently worked on flying tests of radar gear at the Telecommunication Research Establishment, Malvern. He is now Professor of Telecommunication at Imperial College, London. His research interests are in circuit theory, in the theory of communication, and in the psychological aspects of telecommunication.

---

[1] A meta-language is a formal language-system for describing language or a linguistic source. Strictly, a book of grammar is written in a meta-language—it defines a syntax but is itself not literature.



A	1	1	1
B	1	1	0
C	1	0	1
D	1	0	0
E	0	1	1
F	0	1	0
G	0	0	1
H	0	0	0

Figure 2 The simplest form of binary coding of a source of eight independent and equally probable signs.

Before developing the theory, we should perhaps distinguish now between information and meaning. Information has been defined as 'selective potential'—the power to select the right signs from alternatives. But the meaning of these signs arises only when some human users send and receive them, for meaning is not a property of signs, such as words and messages, themselves; rather it is a relationship between signs and their users.

So far we have looked at discrete sources, using separate and distinguishable signs like printed letters. But other sources, such as speech, appear to be continuous. How can these be dealt with? It was the telephone engineer R. V. L. Hartley who answered this question when he published, as early as 1928 [3], the first mathematical theory of communication. The foundations he laid have remained unshaken to this day.

Figure 3(*a*) illustrates a typical example, that of telephone speech. The curve represents the electric current in a microphone, a speech source. It may be considered continuous in time and to vary over all possible amplitudes. But, as Hartley pointed out, any channel of transmission must be regarded as having a finite band-width, *F*. Thus the inset diagram illustrates the long-term spectrum of this signal source. It can be imagined as extending to infinity, but it can be measured and specified only to some limiting but finite frequency *F* cycles per second, as high as we like to make it.

This logical necessity of a finite frequency band enabled Hartley to use what is often called today the 'Sampling Theorem' [8]. Briefly, this states that, within the restriction just mentioned, the waveform of the signal can be specified uniquely by the values of its ordinates spaced apart by equal time intervals, $\frac{1}{2}F$ (figure 3(*b*)). The time origin of these samples is immaterial; only their spacings matter. In other words, the successive ordinates x_1, x_2, . . . could be measured and the values printed as a column of figures; with these figures alone the continuous waveform could be reconstructed by using an 'interpolation function'. But the information of the source is inherent in the discrete samples on their own.

But further, each ordinate, x_r, can be measured only to some finite accuracy. We can specify an ordinate as 2·4 units, or 2·43, or 2·431, but the accuracy is always exactly limited. Quantization is a logical necessity of specification. In this sense, symbols like π, e, $\sqrt{2}$ do not signify magnitudes at all, but rather have the nature of rules. Put another way, they do not represent communicable magnitudes.

Hartley then arbitrarily quantized the amplitudes into discrete intervals, as in figure 3(*b*). In this way, each successive ordinate selects a particular interval, and if we label these intervals with letters of an alphabet, or other discrete signs, the situation becomes identical to that discussed earlier and illustrated by figures 1 and 2. Similarly, we may take over the measure of information rate *H* from equation (1), because, if there are *N* ordinate intervals—the alphabet—each ordinate x_r would require $\log_2 N$ binary digits to specify it as a signal. Thus, as there are $2F$ ordinates per second, the information rate of the source is found [3] to be

$$H = 2F \log_2 N \text{ bits per second} \qquad (2)$$

These aspects have been extensively studied by D. Gabor [6] and D. M. Mackay [7].

Representation of a source in 'phase space'

It is a further result of the Sampling Theorem that the successive waveform ordinates x_r are independent, that is, orthogonal; the specification of any one tells us nothing about the others. It is then very convenient to regard them as co-ordinates of the source $S(t)$. Any segment of the waveform $S(t)$ of duration T then has $2FT$ dimensions and so may be represented as a vector OS in $2FT$-space. Figure 4 illustrates a three-dimensional projection of this, with only three co-ordinates, x_1, x_2, x_3. If other similar, successive, segments of the signal, all of duration T are also represented as vectors in this space, their lengths will all be the same, or nearly the same, provided that T is long. So $2FT$ will then be large, and all signal-points S will lie on the surface of a sphere because such long segments would all be typical of the source.

The analogy to the 'phase-space' of statistical mechanics is evident, with S representing the 'system-point' whose positions represent various states of the system. Further, if our signal-space of figure 4 is quantized into cells (figure 3 (*b*)), we may attach probabilities to these cells. Such probabilities, it will be seen, are probabilities of the various signals of duration T which the source might select—the various 'states' of the source. These in turn depend only upon the probability distribution $p(x)$, the distribution of signal ordinates over the N quantal intervals. Similarly, we may discuss $p(x)$ as the distribution over the signs of an alphabet (figure 1).

Probabilities were introduced into communication theory by Norbert Wiener [4] in his book 'Cybernetics' in 1948. The theory was developed in these terms by C. Shannon [5], showing the close analogies between an information source and a thermodynamic system; we shall look at the theory now in this light.

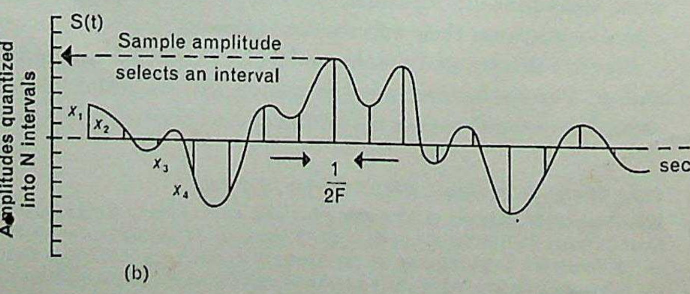


Figure 3 (*a*) A typical continuous source, such as speech. (*b*) The same, quantized on the time-axis and in amplitude.

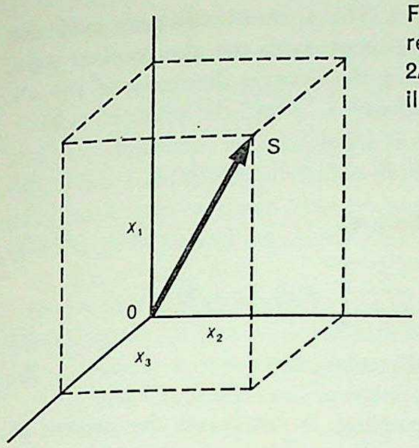


Figure 4 A signal represented as a vector in $2FT$-space ($2FT=3$ illustrated here).

Sign probabilities and information rates—the Wiener-Shannon formula

It was Samuel Morse, or rather his assistant Alfred Vail [9], who first saw intuitively that the less-frequent letters of the alphabet are the more informative. Because, for example, the letters E, T, A occur more often than J, X, Z in English, they can be guessed correctly by the receiver more often. Loosely speaking, the receiver already has prior information that assists him. So Vail planned his code to send more digits of information for the rarer letters than for the frequent ones, so as to specify them more reliably. This is illustrated by his original code (figure 5). He used the number of pieces of type in a printer's office as a measure of the frequency of use of a letter. This original code was not strictly binary, since, unlike the modern version, it used more than simple dots and dashes.

Referring back to the simplest communication model of figure 1, we can attach probabilities $p(x)$ to the signs A, B, ... x. But now, when coding an alphabet, as in figure 2, we can no longer divide it successively into two equal groups, designated 1, 0, such that the total probability of each group is equal. The receipt of a 1 or 0 digit cannot now exactly halve the receiver's doubt. Only in special cases is this possible.

The answer to this dilemma is that, in such circumstances, we should speak only of average information rates, averaged over long times T or over long sequences of signs. Let us, for example, consider a source of print, such as a teleprinter, operating continuously. We may in principle observe each successive long sequence of M letters; let us call this a 'sentence'. In terms of the letter probabilities, $p(x)$, ignoring all statistical constraints between letters, the Central Limit Theorem reveals that all these equally long sentences are equally probable, this probability being P_M, where

$$P_M=\Pi\{p(x)^{Mp(x)}\} \tag{3}$$

We now adopt the idea of a new kind of alphabet or list. Instead of letters A, B, ... x . . . we list all these alternative equally probable sentences. They will naturally total $1/P_M$ in number. Since they are equally probable, their successive binary division into equally probable groups is once again possible. Each such binary division is specified by one binary-digit that halves the receiver's range of doubt. We may now use the logarithmic formula (1) for the average rate of information from this statistically stationary source, writing $1/P_M$ for N (the number of alternatives in the list or ensemble):

$$\begin{aligned} H(x) &= K\log_2(1/P_M) \text{ bits}/M \text{ letters} \\ &= -p(x)\log_2 p(x) \text{ bits/letter} \end{aligned} \tag{4}$$

This is the famed Wiener-Shannon formula for information rate, which is analogous to that for entropy in statistical thermodynamics. Indeed, much has been written about the concept of information rate as 'negative entropy'—of information as being what we need 'to create order out of chaos'.

In (4) we could write more generally the word 'sign' for 'letter', and the same formula is directly applicable to a 'continuous' source, as illustrated by figure 3. There the signs are ordinates $x_1, x_2, \ldots$, and a long sentence has $M=2FT$ signs. The formula (2) again leads to that for $H(x)$ of equation (4).

Redundancy and error-correction

The expression (4) for $H(x)$ has a maximum value when all the signs x are statistically independent and so equally probable. Such an idealized source would have the greatest possible information rate for that particular alphabet of signs. But in practice, the letters of real alphabets are not usually independent. In English, for example, 'TR' is more frequent than 'TU', and Q is always followed by U. Such 'di-gram', and even some 'tri-gram', frequencies have frequently been counted, and statistical constraints of this kind between letters are one way of specifying spelling rules or even syntax rules of language. For purposes of theory, it is necessary to conceive of such, and even longer-range, constraints, though we may have difficulty in attaching precise numerical values.

Such constraints or *a priori* rules reduce the average information rate per sign, but have the advantage that errors in transmission may be corrected by the receiver. For example, this kind of redundancy in English assists us to spot printers' errors.

The redundancy of a source may be increased or decreased by suitably designed codes, and one of the most valuable results of communication theory has been to show that codes may be designed in such a way as to increase a receiver's chance of correcting errors—the design depending essentially upon the statistical properties of the source of errors or 'noise-source'.

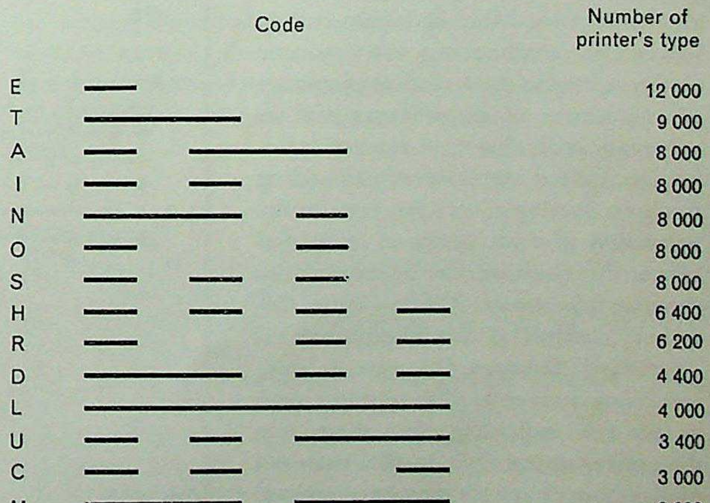


Figure 5 Samuel Morse's (or Vail's) original code, showing longer code-groups for the less frequent letters.

In telephony and telegraphy these disturbances are called noise, and their properties are often well understood; but in other instances, such as in computers, they may not be. Nevertheless, the errors in the received signals are there for the outside observer to see, and so his theory may now embrace two questions:

1. What true information rate is the receiver getting? (Can he find, and so correct, errors himself?)
2. How can the sender encode his signals so as to reduce these errors?

It may surprise the reader to learn that if a receiver is receiving continuous sequences of binary digits, some of which he knows may be wrong, he has a chance of finding out which are wrong and so of correcting them. A simple illustration will have to serve here. Figure 6 illustrates the 'simple parity-check code'. In this diagram, (*a*) shows a sequence of 16 digits in a continuous series. Before these are transmitted, additional digits are built in by arranging the 16 into a 4-row matrix as in (*b*) and then adding in an extra row and column as in (*c*), digits that denote the parity of the whole adjacent row or column. We have written here 1 to signify odd and 0 for even. The term in the corner must always be consistent (0, in this case).

This 25-digit matrix is read off and transmitted. Let us imagine that an error occurs in transmission, and that it is received as the sequence (*d*). The receiver reassembles this as the matrix (*e*) and checks the parity of every row and column; two check-digits are found to be wrong, indicating the coordinates of the erroneous digit (with asterisk in (*e*)). It is important to note that only one error is correctable here in a sequence of 25, though the presence of more could be detected. Since the erroneous digits will usually be distributed randomly, such redundancy improves only the chance of error detection and correction. More elaborate coding will introduce redundancy in a way more closely related to the statistical properties of the source of disturbances and so achieve even higher true rates of information. Indeed, the theory of such coding has been developed very far, even to the correction of short bursts of sequential errors. The question now is, can we raise the true information rate in a noise-disturbed channel of communication indefinitely? Shannon [5] gave a most surprising answer to this problem with a theorem indicating that there is a theoretical upper limit to this rate—no matter how clever we may be at coding, this rate can never be exceeded. Only a sketch of this most important theorem will be given here.

The Shannon capacity theorem

Let us consider the case in which a sent signal x is corrupted by noise and is received as y. What can the receiver do? Since y does not specify each sign selected by the source uniquely, the receiver can now at best assess the relative chances of what might have been sent—that is, he can only infer statistically. Suppose x ranges over the alphabet $x_1, x_2, \ldots x_N$ and that, at some moment in time, a specific signal y_1 be received. This y_1 constitutes only evidence as to which particular x was sent; with this the receiver may compute $p(x/y_1)$, representing the chances that each of the x's might give rise to the evidence y_1. From the product rule

$$p(x,y_1)=p(x)\,p(y_1/x)=p(y_1)\,p(x/y_1) \qquad (5)$$

giving the required distribution:

$$p(x/y_1)=\left[\frac{1}{p(y_1)}\right].\,p(x)\,.\,p(y_1/x)$$
$$=K_1\,.\,p(x)\,.\,L_{y_1}(x) \qquad (6)$$

This equation is most revealing. It interprets the process of communication as the conversion of the receiver's prior state of belief $p(x)$, before receiving any evidence, to a posterior state $p(x/y_1)$, after the evidence. The conversion requires the computation of the likelihood function $L_y(x)$, which is the likelihood that each of the possible x's might give rise to that particular evidence y_1. Such a function is distributed over the whole range $x_1 \ldots x_n$. Let the process of transmission now continue and successive y's be received. Dropping the suffix to y in (5), we may average these expressions over all possible sent x's and received y's. Taking logarithms, we then obtain

$$-\sum_x p(x)\log p(x)+\sum_{x,y} p\,(x,y)\log p(x/y)$$
$$=-\sum_y p(y)\log p(y)+\sum_{x,y} p(x,y)\log p(y/x)$$

Comparing each of these expressions with (4) we see they all

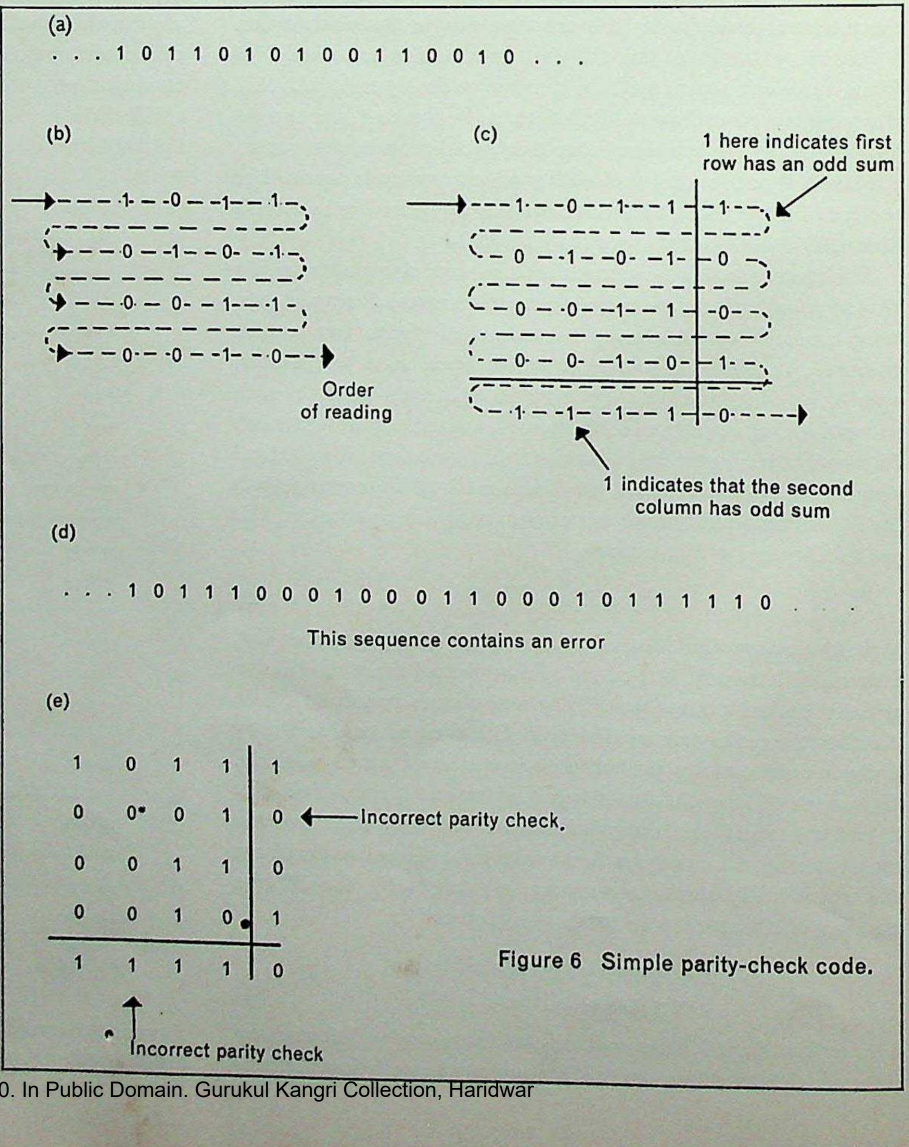


Figure 6 Simple parity-check code.

represent rates of information, with positive or negative signs, so we may express the above equation as

$$H(x)-H(x/y)=H(y)-H(y/x)=R \tag{7}$$

which is Shannon's expression for the true rate of information [5]. Written in this way, the two different ways of expressing this rate R may be given 'common-sense' interpretations if we note the positive and negative signs:

$H(x)$ = Full rate of source.
$H(x/y)$ = Rate of loss of information because of uncertainty about x even when y is received (called by Shannon the 'equivocation', or ambiguity, caused by the noise).

Alternatively:
$H(y)$ = Full rate of the received signals (some of it 'bogus').
$H(y/x)$ = Rate of loss caused by the noise source ('bogus information rate' of noise source).

From this important equation (7) Shannon derived his most interesting Capacity Theorem [5]. Noting that $H(y/x)=H(n)$, the rate of 'bogus information', R, being sent by the noise source, equation (7) may be written

$$R=H(y)-H(n) \tag{8}$$

If the statistical properties of this noise source are given and, as is usually the case, they cannot in any way be altered, then $H(n)$ is fixed; but the received signal y is some combination of sent x and noise n. It is thus conceivable that, by a process of coding the sent signals x, $H(y)$, which is $H(x, n)$, can be maximized. This gives the maximum rate of information possible in this noisy channel:

$$R_{\max}=H_{\max}(x, n)-H(n) \tag{9}$$

$R_{\max}$ is called the channel capacity.

This is of course no proof, nor even clear demonstration, of what is a very remarkable theorem. It means that, given a channel of communication working with a certain source and disturbed by a source of noise, it is possible so to encode the source that error-free messages may be communicated at an information rate approaching, but never exceeding, the capacity, $R_{\max}$. This theorem is as fundamental as the Second Law of Thermodynamics.

Applications

The reader may now ask: what applications has the statistical theory of communication?

The theory has done many things. Most important, it has clarified, for the first time, exactly what it is we are doing when we 'communicate', or receive information. It has shown that such processes are closely analogous to those described by statistical thermodynamics; information reduces uncertainty, and creates order out of chaos. These may be vague phrases, but what matters is that the formula (4) for the information rate of a source corresponds with that for the statistical entropy, with a negative sign, of a thermodynamic system.

Further than this, Shannon's Capacity Theorem, illustrated by equation (9), states that any channel of communication, disturbed by unwanted sources, has a definite information capacity, which is an absolute maximum information rate that it is impossible to exceed, by any way of encoding the sending source. It can be approached, and the price of approaching this ideal rate is a time-delay in transmission. The practical value of being able to calculate this capacity, when we have a communication channel having known properties, is that the efficiency of any method of coding the message source may be stated. Thus the theorem can prevent us trying to do the impossible. Here again is a similarity with thermodynamics. Steam engines were made before the theory had been evolved, but no one knew just how efficient they were.

By defining and distinguishing various concepts—redundancy, information, message, for example—the theory has given guidance to the construction of error-correcting codes. The development of such codes has now reached the most sophisticated levels. Such codes deal not only with simple random noise disturbances but with more realistic kinds, such as 'atmospherics', which may obliterate whole sequences of message symbols.

Communication theory has also made great contributions to solving what is called the detection problem. By this is meant that, when noisy signals are received, the receiver may calculate the likelihood functions L, as in equation (6). However, these merely give the relative probabilities of various messages; they do not say which ones were sent. Some strategy must follow that decides the best choice, in the face of certain penalties for making mistakes. There are two processes then involved in the receiver. The first is to extract all the information out of the noisy signals, resulting in $p(x/y_1)$ in equation (6), and then to make the best decision as to which x should be assumed correct.

These few examples may illustrate the nature of the value of the theory. But, in closing, it should be emphasized again that, in this article, the theory has been sketched (and no more) in terms of the medium of its origin, telecommunication. However, the concepts involved are universal. Applications have been made in many fields of study—to genetics, to psychology of perception, to statistical linguistics, to quantum theory, to physiology of the nervous system, and to scientific method. The theory has made a great contribution to the unity of science.

References

[1] 'Essai sur le 'Cursus Publicus': Sous le Haut-Empire Romain'. Imprimerie Nationale, Paris. 1950.
[2a] 'The world's telephones'. American Telephone and Telegraph Co., New York.
[2b] 'General telephone statistics'. International Telecommunication Union, Geneva, and other sources.
[3] Hartley, R. V. L. *Bell. Syst. tech. J.*, **7**, 535, 1928.
[4] Wiener, Norbert. 'Cybernetics'. Technical Press of M.I.T. and John Wiley & Sons, New York. 1948.
[5] Shannon, C. *Bell. Syst. tech. J.*, **27**, 379, 623, 1948.
Idem. 'The mathematical theory of communication'. University of Illinois Press, Urbana. 1949.
[6] Gabor, D. *J. Instn Elec. Engrs.*, **93**, Pt. III, 429, 1946.
[7] Mackay, D. M. *Phil. Mag.*, **41**, 289, 1950.
[8] Whittaker, J. M. 'Interpolatory function theory'. Cambridge University Press, London. 1915.
[9] Garratt, G. R. M. *New Scientist*, **16**, 393, 1962.
[10] Fisher, Sir Ronald. 'The Design of Experiments'. Oliver and Boyd Ltd, London. 1935.

William Herschel and sidereal astronomy

Michael Hoskin

Sir William Herschel is famous for discovering the first 'new' planet in recorded history, and he was also an assiduous observer of stars and compiler of star-catalogues. These achievements were recognized by his contemporaries, but, nevertheless, his major papers describing his theories of the structure of the heavens had little influence in his day. This article discusses the assumptions on which he based his theories and the reasons for their being generally ignored during his life.

William Herschel, it is often said, was the father of sidereal astronomy. Nearly always in the history of science it is futile, if not positively harmful, to ascribe paternities, for science develops in a highly complex fashion, and no major field is ever the creation *ex nihilo* of one man. But Herschel has a special place in the history of sidereal astronomy, for although he was not the first to theorize about the fixed stars as more than mere reference points against which to plot the movements of the planets, he was the first to base his theories on the evidence of the great telescopes without which a science of sidereal astronomy is impossible.

Herschel was born in Hanover in 1738, the son of Isaac Herschel, a musician. After the occupation of his city by the French in 1757, William and his brother Jacob fled to England, and by 1767 he was comfortably established in Bath as organist of the Octagon Chapel. By this time, his insatiable curiosity had taken him from the mathematical theory of harmony to geometrical optics; from optics to the construction of telescopes; and from telescopes to astronomy. Herschel read in popular works that the distances of the stars were unknown, because they were so great that the parallax of the stars caused by the Earth's motion had escaped detection. He also learned of the importance of wide apertures in telescopes intended to grasp the light of faint objects. He quickly decided that he must tackle this fundamental problem of stellar distances, but soon he found that he would have to build for himself the very large telescopes he required. In 1773 he heard of a Bath neighbour who had made a hobby of grinding and polishing mirrors for telescopes, and Herschel bought up his stock of tools, equipment, and unfinished mirrors. At that time, telescopic mirrors were made of speculum metal, and Herschel began by having the disks cast for him and then undertaking the grinding, figuring, and polishing himself. He learned what he could from Robert Smith's 'Compleat System of Opticks'; the rest he discovered for himself as he went along. 'To my sorrow', wrote his sister Caroline, 'I saw almost every room turned into a workshop . . . I was even obliged to feed him by putting the Vitals by bits into his mouth;—this was once the case when at the finishing of a 7-feet mirror he had not left his hands from it for 16 hours together.'

Herschel's instruments increased rapidly in focal length, for greater magnification: 5½ feet, 7 feet, 10 feet, even 20 feet. By 1781 he was planning a monster telescope with a focal length of 30 feet and an aperture of 48 inches. The early months of the year saw a series of experiments to find a suitable metal for such a large mirror, and when these failed he decided to compromise with a 36-inch mirror of speculum metal. But even this proved too much for the local foundries. Herschel responded characteristically to this setback by building his own foundry in the basement of his house in Bath. Some spectacular failures with the molten metal, however, convinced him that for the present, at least, he had overreached himself.

Meanwhile he had to contine with the busy musical practice on which his livelihood depended, but still he found time to make systematic reviews of the entire visible sky. It was on 13 March 1781, while engaged in the second of these reviews, that he noticed a star that seemed to show the visible disk that is characteristic of comets and planets. Sure enough, a few days later the 'star' had moved position. Not suspecting for a moment that he had found the first primary planet of the solar system to be discovered since recorded history began, he wrote up his observations under the title 'Account of a comet', and a friend duly forwarded his paper to the Royal Society.

A storm of controversy immediately broke about Herschel's head. It was not enough that he had shown incomparable skill by recognizing at a glance the unusual nature of an object that professional astronomers could rediscover only with great difficulty; he had offended by mentioning casually, though truthfully, that he was using magnifications of up to six thousand times. To make such claims, scoffed his incredulous critics, he must be 'fit for Bedlam'. In the end, it was only by taking a telescope to Greenwich and setting it up alongside the best of the Royal Observatory instruments that Herschel vindicated himself. But when it was realized that the 'comet' was actually the planet now known as Uranus, Herschel

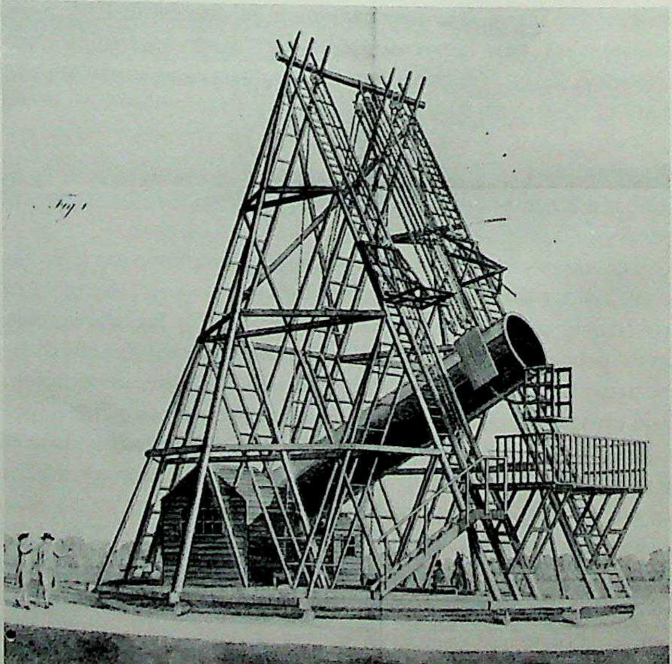


Figure 1 Herschel's greatest telescope. The focal length is forty feet. (All illustrations are by courtesy of the editor of the Philosophical Transactions of the Royal Society.)

M. A. Hoskin, *Ph.D.*

Was born in London in 1930 and was educated at London University and at Cambridge University, where he was Research Fellow in mathematics at Jesus College. He is now Lecturer in History of Science at Cambridge. His main interests are in the history of astronomy, and he is now working on theories of the Galaxy and nebulae in the century following Herschel's death in 1822.

became famous overnight. And although the discovery was no more than a by-product of his study of the fixed stars, it proved to be of crucial importance, for it earned him a royal pension. Now at last, in middle age, he could give up music and devote all his great energies to astronomy.

For the rest of his life, Herschel lived within a few miles of Windsor Castle, and it was at Slough that he died in 1822. During these forty years he was able to give the whole of his time to his beloved stars, the even tenor of his work scarcely interrupted even by family events such as his marriage in 1788 and the birth of his son John in 1792. At first sight the story is of one long succession of achievements. He built a gigantic telescope of forty foot focal length with a mirror weighing a ton (figure 1). He compiled monumental catalogues of double stars, of star clusters and nebulae, and of the comparative brightnesses of stars, and so laid the foundations of what he termed the natural history of the heavens. And he published many papers on the Sun, the Moon, and the planets and comets of the solar system, and on his discoveries of infra-red rays and of the existence of binary stars. But, as he himself said, 'a knowledge of the construction of the heavens (that is, of the large-scale structure and development of stars and nebulae) has always been the ultimate object of my observations', and it is when we penetrate beneath the surface of his lengthy papers on the construction of the heavens that we uncover a fascinating mosaic of brilliant insights, bitter disappointments, and at times a wilful refusal to face the implications of new evidence.

Herschel's first and fundamental problem, that of stellar distances, was to haunt him throughout his career. It was clear from the start, firstly, that accurate measurements of the distances of the nearer stars would be possible only when their annual parallaxes had been measured, and, secondly, that the best hope of making so delicate a measurement lay in observing the movement of the near star relative to a second star sufficiently distant to be, for all practical purposes, fixed. Herschel accordingly catalogued all the 'double stars' he could find, in the belief that he would find in his catalogues many suitable pairs. But in the meantime he decided to measure the relative brightnesses of stars and to interpret these to give their relative distances from the Earth; for it was obvious to him that the fainter stars are, by and large, further away than the brighter ones: the same star removed to twice its present distance will have its brightness reduced by three-quarters.

This line of argument raised two problems for Herschel. First of all, exactly how far could this generalization be carried? And secondly, how were the apparent brightnesses to be measured? Herschel habitually thought in biological metaphors, and he answered the first question by allowing the stars only as much variation as occurs among an animal species. We now know that two stars can differ in absolute brightness in the proportion of millions to one, but Herschel would permit only very small, indeed negligible, variations. He answered the second question, in default of adequate techniques, by assuming that the age-old traditional magnitudes directly represented distances, so that a faint star listed as of the sixth magnitude would be six times further than—or, more exactly, one-thirty-sixth the brightness of—a prominent star of the first magnitude.

These two postulates, and Herschel recognized them as such, were essential to his programme of work; destroy them, and the task to which he had dedicated his life was hopeless. He was therefore unable to take a disinterested view of criticisms of them. And so, when in 1782 they were challenged by the Astronomer Royal, Nevil Maskelyne, as chairman of the Royal Society Committee on Papers, Herschel dug in his heels. In particular, Maskelyne pointed out that according to a paper published in 1767 by John Michell the drop in apparent magnitude from first to sixth corresponded to an increase in distance by a factor of about 25, and not 6 as Herschel postulated. Herschel replied by defining new apparent magnitudes so as to give the distance factor he required. This was of course a good defensive step, but one of no positive value. As long as he could not measure brightness, the traditional magnitudes were all he had to use.

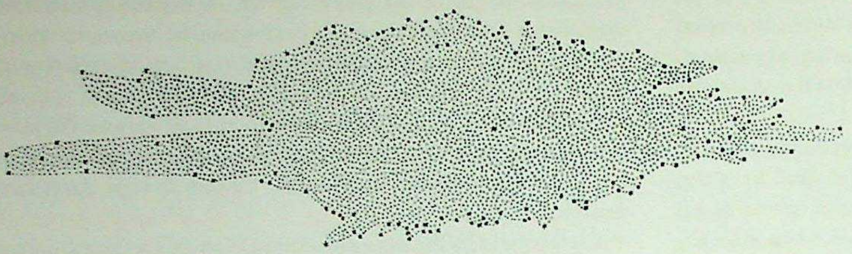
Characteristically, for many years Herschel avoided carrying through a simple test that would have exposed the postulate as untenable, namely, that of counting the total numbers of stars of each successive magnitude and examining whether the figures were reasonable when translated into the alleged distances. Several times he considered this test, while other astronomers grew more and more sceptical about the validity of the postulate. In 1801, Olbers made measurements that suggested the figure 10 in place of Herschel's 6, and not long afterwards Herschel admitted 'some doubt' about the postulate after the first couple of magnitudes. But as soon as he required it for some calculations, he did not scruple to suppose it accurate to three significant figures.

It was only in the last decade of his life that Herschel succeeded in comparing the light of two stars by taking two similar telescopes, pointing one at each of the stars under comparison, and restricting the aperture of the telescope pointing at the brighter star until the stars appeared as equal in brightness. The ratio of their apparent brightnesses could then be calculated from the apertures of the telescopes. Herschel found that his figure of 6 should have been nearer 12—and only then applied the test of star counts to give further proof of the falsity of a postulate that had now outlived its usefulness.

His other postulate, that the stars vary little in themselves, so that brightness (properly measured) is an index of nearness, was more fundamental. When he began his career, the evidence against this postulate consisted partly in the observed variety of colour among the stars, and, partly, by an argument from analogy, in the great variety to be found in the solar system itself. There was one other argument of importance, due again to the remarkable paper by John Michell. Michell had been struck by the fact that the stars occur in groups far more often than coincidence would allow. It was next to a certainty, Michell argued, that the grouping of the Pleiades is not simply an optical effect due to the chance position of us the observers. They must form a physical system of stars grouped together in space. In the same way, he argued, double stars occur far too often to be explained as pairs of stars that happen to lie in the same direction from Earth. Many of the stars in these pairs must be close to one another in space.

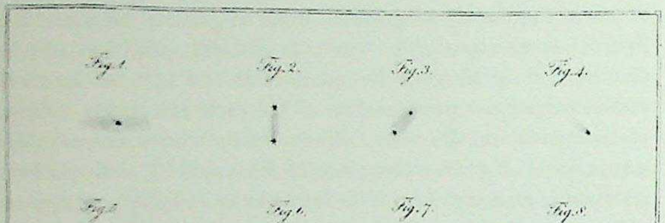
Herschel chose to ignore this threat to his attempt to measure parallax, which presupposed that double stars were not close together in space. He had more justification than usual for taking this attitude, for Michell's argument applied only to double stars in general, and not of course to every double star. Herschel, on the other hand, was searching for particular 'optical' doubles, and he was well aware that Michell's arguments were less forceful when applied to areas of sky thickly strewn with stars. But Herschel was emphatically not justified in ignoring the implied threat to his uniformity postulate: the Pleiades contain stars of widely differing apparent magnitudes, and Michell had shown that it was nearly certain that they were all at about the same distance from us.

This situation became still more serious in 1802, when, as Michell had predicted, Herschel found that several of his double stars were binary systems revolving about a common centre. Among his confirmed binary systems were two whose



2

Figure 2 Cross-section of the Milky Way according to Herschel.



3

Figure 3 Sketches illustrating Herschel's later theory of the development of nebulosity under attractive forces.

members he describes as 'very unequal', and he now knew for certain that these stars were at roughly the same distance. In other words, from this time on he had positive visual evidence that the variety that exists among the stars is far greater than he had supposed, and yet he still continued to suppose an almost perfect uniformity whenever his purposes required it. As for other astronomers, their doubts grew into a certainty when in 1812 Bessel found that the faint star 61 Cygni had an enormously large proper motion, moving across the sky at over 5 seconds per annum. Surely, they argued, this movement must mean that the star is very near, notwithstanding its apparent faintness. As a result, proper motions quickly superseded apparent magnitudes as the rough index of distance.

Similar problems of distance measurement undermined Herschel's achievements in the study of other aspects of the construction of the heavens. In 1783, through the analysis of a small and already outdated list of proper motions of stars, he succeeded in detecting a pattern which, he believed, reflected the motion of the solar system through space. He argued that just as the daily spin of the Earth appears to us as a daily motion of the entire sky, so the motion of the solar system through space will appear as an opening out of certain stars in front of us and a closing in of others behind. The direction he derived is very close to the best modern value, an astonishing achievement that owed nothing whatever to his superiority as an observer. But when he reworked the problem twenty years later on the basis of fuller information, he ran into difficulties because of his slavish adherence to brightness as the sign of nearness, and he extricated himself only by playing an involved confidence trick on his readers. Had the motion of 61 Cygni been announced at this time, he would presumably have been forced to abandon this search for the motion of the solar system.

Another fundamental question to which distance measurements proved crucial was that of the nature of nebulae. Herschel's interest in them was aroused when a friend presented him with a catalogue of nebulae and star clusters compiled by the French comet-hunter Charles Messier. Messier believed that many of the nebulae in the catalogue were not composed of stars at all, although it was of course obvious that a distant star cluster in which individual stars could not be detected would appear nebulous. Was Messier right?

Now Herschel had the most powerful telescopes in the world, and he soon found that some of Messier's nebulae were certainly star clusters. Others he thought, not always rightly, would be seen to be star clusters as soon as a more powerful telescope was used. Why then postulate the existence of a mysterious self-luminous fluid when all the evidence could be explained in terms of star clusters alone? The more Herschel thought of it the more convinced he became, and in no time at all he was writing to a friend that he had resolved into stars 'almost all' the nebulae of Messier's list.

So far, so good. Meanwhile, Herschel had set out to make a three-dimensional map of the Milky Way. Like Wright, Kant, and Lambert, he believed that the milkiness is an optical effect due to our immersion in the middle of a layer of stars. If we look out sideways we see few stars, and those near, before we come to empty space; but if we look along the layer of stars then we see stars and yet more stars, and these appear to us as a milkiness.

Herschel saw that he could actually trace the outline of this layer of stars if he made two assumptions. Firstly, that his telescopes could penetrate to the borders of the Milky Way in all directions; and secondly, that the stars in this system of ours are spread out uniformly, so that when we see more stars in some directions than in others we are seeing more space stocked with stars and not more stars packed into the same space.

In later years, when he found that even his great 40-foot reflector still showed the glimmer of barely visible stars, he realized that he could not after all penetrate to the borders in all directions, and as he greatly extended Messier's catalogue of nebulae and star clusters, he came to appreciate that the distribution of stars is very far from uniform. But for the present all seemed to be well. He had only to point his telescope in a given direction, count the number of stars he could see, and apply a simple formula, and he arrived at the distance to the border of the Milky Way in that direction. He called the method 'star-gauging'.

R
505
L84E
V. 23
~~63443~~

Herschel counted and calculated, turning a blind eye to instances where the results suggested that faint stars were near, and in due course was in a position to draw his famous map of a cross-section of the Milky Way (figure 2). This done, he could turn his attention again to the unresolved nebulae. These he believed to be distant star systems; but some of them, like the great nebula in Orion, covered large areas of sky, and to be so far and yet appear so large such nebulae must be enormous in size. In fact, concluded Herschel, they 'may well outvie our Milky Way in grandeur'.

Herschel admitted himself puzzled by what he called 'planetary nebulae', nebulae that presented the round disk-like appearance of planets. As the camera and spectroscope were later to show, these nebulae are spheres or shells of gas surrounding a central star often too faint to be seen visually. It was in 1790 that Herschel happened upon a planetary nebula in which the central star was clearly to be seen. 'The star is perfectly in the centre', he noted, 'and the atmosphere is so diluted, faint, and equal throughout, that there can be no surmise of its consisting of stars; nor can there be any doubt of the evident connection between the atmosphere and the star.' He had been too hasty in supposing 'that all milky nebulosity, of which there is so much in the heavens, is owing to starlight only'.

The consequences were clear. First of all, his rule for judging the distance of nebulae must be abandoned. 'Unresolved nebulosity' might be a distant star system; but there was no guarantee of this, and no guarantee therefore of the existence of galaxies comparable to the Milky Way. This, coupled with his realization that his mapping of the Milky Way was based on untenable assumptions and that in some directions the Milky Way was 'fathomless', forced him to reconsider his earlier 'island universe' theory. Robbed of his distance criterion, he had no way of telling whether there were star systems independent of the Milky Way. All that he could say for sure was that, for him, it was 'the most brilliant, and beyond all comparison the most extensive sidereal system'.

The second, and happier, consequence, was Herschel's immediate realization that 'if, therefore, this matter is self-luminous, it seems more fit to produce a star by its condensation than to depend on the star for its existence'. Years before, in his attempts to make sense of the thousands of nebulae and star clusters he had discovered, he had picked on the degree of clustering as a main characteristic, and he had gone on to interpret this as an index to the age of the nebula or cluster. The more condensed the cluster, it seemed to him, the longer had attractive forces been operating in it. He had also been struck by the uneven distribution of nebulae across the sky and thought they lay in strata, of which the Milky Way was an example of one before fragmentation. In 1790 he was therefore already long accustomed to thinking of what we would term the 'evolution' of star systems. The recognition of 'true nebulosity' meant that he must reduce the emphasis on overall spatial distribution, which was now shrouded in uncertainty, and instead carry his theory of development in time back to the stage before individual stars were formed out of the self-luminous matter.

Two monumental papers in which he expounds this theory were published in 1811 and 1814, when he was well into his seventies. He inferred the history of the apparently unchanging stars by arranging items from his catalogues in 'articles', each article designed to represent a stage in development (figure 3). 'It will be found that those contained in one article, are so closely allied to those in the next, that there is perhaps not so much difference between them, if I may use the comparison, as there would be in an annual description of the human figure, were it given from the birth of a child till he comes to be a man in his prime.'

Several of the articles are planned to present the formation of a single star, and it is a measure of the damage wrought by the collapse of his criterion of distance for nebulae and star clusters that, in the majority of his examples, these alleged stars-in-the-making are actually entire galaxies external to the Milky Way. Even the great Andromeda nebula, nearest of the galaxies comparable to the Milky Way system, he now regards as 'ambiguous': he is unsure whether it is stellar in nature or not.

A century after Herschel's death many of these questions were still being hotly debated. The size of the Milky Way system was still quite uncertain, and it was disputed whether the Andromeda and similar nebulae were associated with it or were separate galaxies in their own right. It is not surprising, then, that Herschel was unable to reach a definitive solution to these problems. But when we study the writings of Herschel's contemporaries, we find that, although they recognize his pre-eminence as an astronomer, they are little influenced by his great papers on the construction of the heavens, and this lack of influence calls for an explanation.

The answer lies partly in incidental factors. Thus for many years it was thought, wrongly, that the solar motion he had measured was illusory. Also astronomers were unaccustomed to inferring change from objects that appeared unchanging, and they insisted on waiting until Herschel's alleged changes had actually been observed taking place. But there were, as we have seen, fundamental structural weaknesses in Herschel's theories. These were the assumptions, especially of uniformity of size and distribution among the stars, that were, to other astronomers, at best dubious and sometimes in direct conflict with the evidence.

Bibliography

'The Scientific Papers of Sir William Herschel', edited by J. L. E. Dreyer. The Royal Society and the Royal Astronomical Society, London. 1912.

Armitage, A. 'William Herschel'. Nelson, London. 1962.

Buttmann, G. 'Wilhelm Herschel: Leben und Werk'. Wissenschaftliche Verlagsgesellschaft, Stuttgart. 1961.

Hoskin, M. A. 'William Herschel: Pioneer of Sidereal Astronomy'. Sheed & Ward, London. 1959.

Idem. 'William Herschel and the Construction of the Heavens'. Oldbourne Press, London. 1963.

Lubbock, Constance A. 'The Herschel Chronicle'. Cambridge University Press, Cambridge. 1933.

Sidgwick, J. B. 'William Herschel: Explorer of the Heavens'. Faber and Faber, London. 1953.

077815
गुरुकुल कांगड़ी विश्वविद्यालय हरिद्वार

Modern thermochemistry

C. T. Mortimer and H. D. Springall

Great advances have been made in the techniques of thermochemistry during the last twenty-five years. The heats of combustion of a wide variety of compounds can now be determined very accurately, using greatly improved static methods and the new moving-bomb calorimetry. Measurements of heats of reaction, such as hydrogenation and hydrolysis, can yield valuable information about the structure and stability of molecules. Elsewhere in this issue N. Bartlett discusses the significance of thermochemistry in the study of the noble gases.

The aims of thermochemistry

Thermochemistry is the study of the changes of heat content, or enthalpy, ΔH, in chemical reactions. Some of these heat changes have been found from spectroscopic and electron-impact experiments, or from kinetic and equilibrium data, but the majority have been measured calorimetrically. The heats of many reactions are of interest in themselves, but they also provide a means of calculating the heat of formation of a compound, $-\Delta H_f^\circ$. This is the heat liberated when a compound is formed in its standard state from its elements in their standard states, that is, the states in which they normally exist at 298°K and atmospheric pressure. The superscript ° indicates that reactants and products are in their standard states. The heat of formation is often not directly measurable, but it can be derived from the heat of combustion. The procedure is indicated in the enthalpy-level diagram (figure 1), where the heat of formation of an organic compound of the general formula $C_\alpha H_\beta O_\gamma N_\delta$, which is liquid at 298°K and atmospheric pressure, is seen to be the difference between the sums of the known heats of formation, $\Sigma[-\Delta H_f^\circ$ (products)], of the products of combustion, namely gaseous carbon dioxide and liquid water, and the measured heat of combustion, $-\Delta H_c^\circ$, of the liquid compound.

The subject had its origin in the monumental work of M. Berthelot and of J. Thomsen at the end of the last century, when it was thought that the enthalpy change was the sole driving force of a chemical reaction. With the realization that it is actually the corresponding Gibbs free-energy change, ΔG, given by $\Delta G = \Delta H - T\Delta S$, that is the driving force, interest in thermochemistry waned because of the inaccessibility of the values of the entropy changes, ΔS. Interest revived in the 1930s with the development of molecular structural studies. The revival was largely due to two factors. Firstly, it was realized that although it would be preferable, for molecular studies, to use the more fundamental Gibbs free-energy changes and to correct to 0°K and zero pressure, nevertheless the inaccessible conversion factors are small. Consequently, the experimentally available ΔH_{298} terms can be used without introducing substantial error. Secondly, the work of F. D. Rossini greatly improved the accuracy of calorimetric methods; he insisted that they be applied only to highly purified compounds. He defined [1] one of the main aims of the thermochemist as the building up of a comprehensive set of heats of formation for a wide variety of compounds, so that the heat of any reaction may be calculated. Such knowledge would be of great practical value. It would permit, for example, the forecasting of the most suitable conditions for a new industrial process, or the calculation of the potential thrust of an untried rocket fuel.

C. T. Mortimer, *M.Sc., Ph.D., F.R.I.C.*

Was born in 1927 and was educated at Ackworth School and Manchester University. After a short period at the University of Khartoum he became a Lecturer (subsequently Senior Lecturer) in Chemistry at the University of Keele, where he has been engaged in thermochemical research for the last eight years. During this time he has visited the University of Ottawa to carry out research in microcalorimetry, and the Thermochemical Laboratories at the University of Lund.

H. D. Springall, *M.A., D.Phil., F.R.I.C.*

Was born in 1910 and was educated at Colfe's Grammar School, London, and Lincoln College, Oxford, where, as a pupil of N. V. Sidgwick, he developed an interest in molecular structural problems. He was a Commonwealth Fund Fellow in the California Institute of Technology with Professor L. Pauling from 1936 to 1938. During 1938–1939 he returned to Oxford, working with Sir Robert Robinson on problems of protein structure. From 1939 to 1945 he was in the Armaments Research Department of the Ministry of Supply. From 1945 to 1950 he was a Lecturer (subsequently Senior Lecturer) in Chemistry in the University of Manchester. Since 1955 he has been Professor of Chemistry in the University of Keele.

More fundamentally, a study of heats of reaction (especially of formation) can provide a wealth of information about the strengths of chemical bonds. The $-\Delta H_f^\circ$ term for a compound can be combined with latent heat terms and with the heats of atomization of the elements from their standard states, to give the heat of formation of the compound, in the gas phase, from the free ground-state atoms of its constituent elements. This is known as the atomic heat of formation, $-\Delta H_{f,a}$; it is an important molecular constant that can be analysed into generally applicable mean bond-energy terms $\bar{E}$(b).

For a simple molecule, such as methane, $-\Delta H_{f,a} = \Sigma\bar{E}(b)$. For molecules for which more than one valence-bond structure can be written, such as benzene, $-\Delta H_{f,a}$ is greater than $\Sigma\bar{E}$(b) by an amount E_R, known as the resonance, or stabilization, energy (for benzene, $E_R = 36$ kcal/mole):

$$-\Delta H_{f,a} = \Sigma\bar{E}(b) + E_R$$

The thermochemical demonstration [2*a*] that the E_R term for cyclo-octatetraene was as low as 5 kcal/mole contributed to the realization that the compound was not aromatic. For molecules that are strained, such as cyclobutane derivatives, $-\Delta H_{f,a}$ is less than $\Sigma\bar{E}$(b) by an amount E_S, which is the strain energy due to distortion of intervalency angles and repulsion forces between non-bonded atoms:

$$-\Delta H_{f,a} = \Sigma\bar{E}(b) - E_S$$

Thus the closing of the cyclobutane ring in the formation of biphenylene is accompanied by a strain, or destabilization, energy of 64 kcal/mole [2*b*].

Discussions of the derivation of bond strengths from heats of reaction for a wide variety of compounds—including organometallic compounds and co-ordination complexes—and extensive sets of bond-energy data have been given by, among others, T. L. Cottrell [3] and C. T. Mortimer [4].

Methods of combustion calorimetry

Static bomb. The modern static-bomb calorimeter, developed from the classical Berthelot-Thomsen steel bomb, retains an important place in experimental thermochemistry. Three major developments are the use of very accurate platinum-

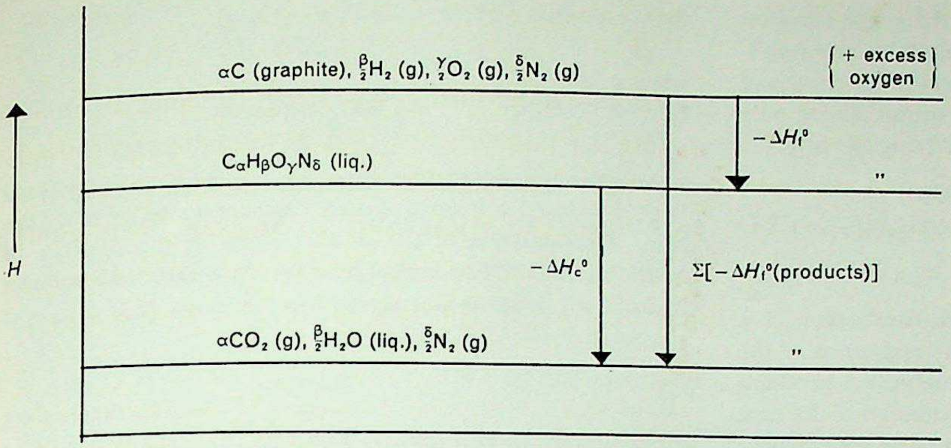
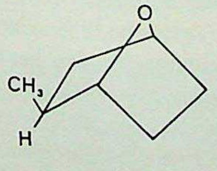
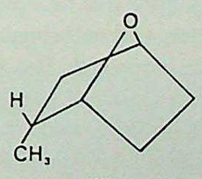

αC (graphite), $\frac{\beta}{2}H_2$ (g), $\frac{\gamma}{2}O_2$ (g), $\frac{\delta}{2}N_2$ (g)
+ excess oxygen
$-\Delta H_f°$
$C_\alpha H_\beta O_\gamma N_\delta$ (liq.)
"
H
$-\Delta H_c°$
$\Sigma[-\Delta H_f°$(products)]
αCO₂ (g), $\frac{\beta}{2}H_2O$ (liq.), $\frac{\delta}{2}N_2$ (g)
"

Figure 1 Energy-level diagram showing relation between heat of formation from standard-state elements, $-\Delta H_f°$, and heat of combustion, $-\Delta H_c°$, for a compound $C_\alpha H_\beta O_\gamma N_\delta$.

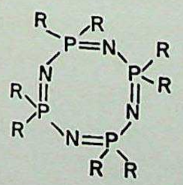
resistance thermometry; the use of corrosion-resistant nickel alloys, instead of steel, for the construction of the bomb; and the twin-valve bomb-head introduced by E. J. Prosen and F. D. Rossini in 1944 [5]. The twin-valve head enables the gaseous products of combustion to be flushed out of the bomb, permitting estimation of the carbon dioxide produced in the combustion of organic compounds and hence calculation of the precise fraction of the sample (usually more than 99·9 per cent) that has actually burnt. This type of calorimeter—which yields, initially, constant-volume intrinsic-energy (ΔU) data—can give heats of combustion to an accuracy of ± 0·01 per cent when applied to compounds containing only carbon, hydrogen, oxygen, and nitrogen. For example, the heat of transformation from the *exo* to the *endo* form for the two isomers of 2-methyl-7-oxa-bicyclo[2,2,1]heptane (I and II) has been determined [6] from the difference between their measured heats of combustion, $-\Delta H_c° = 1007{\cdot}58 \pm 0{\cdot}18$ and $1008{\cdot}43 \pm 0{\cdot}14$ kcal/mole respectively; this corresponds to $\Delta H = 0{\cdot}85 \pm 0{\cdot}23$ kcal/mole. Knowledge of such heats of isomerization should ultimately yield valuable information on the repulsion forces operating between non-bonded atoms.



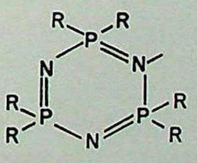
(i)



(ii)



(iii)



(iv)

In addition, exploratory studies have been made on compounds containing elements other than those listed above. Moderately accurate heats of combustion have been obtained [7] for the phosphonitrilic derivatives (III) and (IV), in which $R = CH_3$ or C_6H_5. These indicate that the cyclic, conjugated systems in these compounds are considerably stabilized, due to delocalization of 3*d*-electrons of the phosphorus atoms.

Rotating bomb. The static-bomb calorimeter can give really accurate data only for compounds containing only carbon, hydrogen, oxygen, and nitrogen. The new technique of rotating-bomb calorimetry [8], which results in the mixing of the contents of the bomb after combustion, has made it possible to measure the heat of combustion of a wide range of organic compounds, containing such elements as sulphur, the halogens, and even metals, as well as inorganic compounds, for which very precise data can now be obtained.

The rotating-bomb calorimeter, shown in figures 2 and 3, in which the cylindrical bomb is made to rotate simultaneously about its own long axis and one perpendicular to it, half-way along the bomb, was designed by S. Sunner of the University of Lund. This apparatus is being used by us to determine the heats of formation of organometallic and organometalloid compounds.

Figure 3 shows the bomb and its lid, which is held in place by the massive screw cap. A platinum wire frame, fastened to the lid, supports the crucible that holds the sample to be burnt. When the bomb is rotated the crucible falls from the support and is washed by the aqueous solution in the bomb.

The importance of this development in bomb calorimetry is evident from consideration of the requirements that have to be met to obtain accurate heats of combustion; precise analysis of the combustion products is quite as important as the measurement of heat change. In the combustion of compounds containing only carbon, hydrogen, and oxygen, the final state after the reaction is well defined, but difficulties arise if other elements are present. Thus, when an organic sulphur compound is burnt in a static bomb, the concentration of the resulting sulphuric acid may not be uniform throughout the bomb, and the question arises of the average overall concentration of sulphuric acid to be used in the thermochemical calculation. The question is important because of the high heat of dilution of sulphuric acid. This problem is obviated by rotating the bomb after combustion so that the final solution of sulphuric acid is homogeneous.

A different kind of difficulty, common to both static and moving bombs, arises in combustions in which one of the products is a solid; for example, a metal oxide resulting from the combustion of an organometallic compound. Here there is uncertainty about the crystalline form, particle size, and degree of hydration of the solid product, all of which can slightly affect its heat of formation. These uncertainties can be removed in an elegant way in the rotating bomb by putting in it some aqueous solution in which the solids will dissolve. For example, W. D. Good, D. Fairbrother, and G. Waddington [9], when measuring the heat of combustion of dimanganese decacarbonyl, placed a dilute solution of nitric acid in the bomb; the only reaction products were gaseous carbon dioxide and a solution of manganous nitrate:

$$Mn_2(CO)_{10} + 4HNO_3 + 6O_2 \rightarrow 2Mn(NO_3)_2 + 10CO_2 + 2H_2O$$

Good [10] has developed the idea a stage further for organo-silicon compounds by ensuring that the silicon is both oxidized and rendered into a water-soluble form in the actual combustion process. Thus, the heat of combustion of hexamethyldisiloxane was measured by burning a mixture of the siloxane and trifluoromethylbenzene in 1 : 4 molar proportion, giving only carbon dioxide and an aqueous solution of fluosilicic acid:

$$(CH_3)_3Si{-}O{-}Si(CH_3)_3 + 4C_6H_5CF_3 + 42O_2 \rightarrow 34CO_2 + 2H_2SiF_6 + 17H_2O$$

Ancillary thermochemical data

Even with the straightforward static-bomb combustion of a simple organic compound, once the heat of combustion of the compound has been measured, and the combustion products have been identified and estimated, the calculation of the heat of formation depends on the heats of formation of the products being known. There are comparatively few of these 'key' products, and the accuracy of a large section of thermochemical data depends on the precision with which their heats of formation are known.

Carbon dioxide and water are the most important of these key substances, because they are the main (often sole) combustion products in nearly all static-bomb combustions, and they frequently arise also in rotating-bomb combustions. Fortunately, the heats of formation, $-\Delta H_f°$, of gaseous carbon dioxide and of liquid water are now known with very great accuracy: the values are $94{\cdot}052 \pm 0{\cdot}011$ and $68{\cdot}315 \pm 0{\cdot}101$ kcal/mole respectively. They are based on the experiments of Rossini and his collaborators in 1945 [11], and corrected using the 1961 atomic weights based on ^{12}C.

In the combustion of nitrogenous organic compounds, in either the static or moving bomb, most of the nitrogen separates as gaseous nitrogen. A small proportion of it, however, is oxidized, and this is isolated and estimated as aqueous nitric acid; the heat of formation of this is needed so that a correction can be made.

With the extension of combustion calorimetry to more varied types of compounds in the moving bomb, the appropriate thermochemical characteristics—such as heats of formation, solution, and dilution—of all additional reagents and novel products must be either known or determined. Typical new combustion products arising in these reactions are metal oxides and aqueous solutions of sulphuric, phosphoric, hydrofluoric, and fluosilicic acids.

Fluorine has been used as an oxidant in studies of substances not amenable to combustion in oxygen, and the heats of formation of a number of metal and non-metal fluorides have been determined. As the thermochemistry of fluorine compounds develops and the use of fluorine as an oxidant finds wider application to organometallic compounds, these fluorides are likely to become important key substances. Experimentally, the problem of corrosion is very serious with both fluorine and hydrofluoric acid.

The determination of the heat of formation of crystalline silica (α-quartz) affords an interesting example of the versatility of modern combustion methods. Some years ago, G. L. Humphrey and E. C. King [12] measured the heat of combustion of silicon in oxygen, using a static-bomb calorimeter. However, the resulting mixture contained unburnt silicon, and as this was difficult to estimate, the calculated value of $-\Delta H_f°$ (SiO_2, α-quartz) $= 209{\cdot}9 \pm 1{\cdot}0$ kcal/mole was not very reliable. Using new techniques, J. L. Margrave and his co-workers [13] have measured the heats of combustion of silicon and α-quartz in gaseous fluorine:

$$Si(\text{cryst.}) + 2F_2(g) \rightarrow SiF_4(g) \qquad -\Delta H = 386{\cdot}02 \pm 0{\cdot}24 \text{ kcal/mole}$$

$$SiO_2(\alpha\text{-quartz}) + 2F_2(g) \rightarrow SiF_4(g) + O_2(g) \qquad -\Delta H = 168{\cdot}27 \pm 0{\cdot}24$$

whence for

$$Si(\text{cryst.}) + O_2(g) \rightarrow SiO_2(\alpha\text{-quartz}) \qquad -\Delta H_f° = 217{\cdot}75 \pm 0{\cdot}32$$

An entirely independent approach has been made by Good [14], who used a rotating-bomb calorimeter to measure the heat of combustion of silicon in oxygen in the presence of hydrofluoric acid. The heat of reaction of α-quartz with aqueous hydrofluoric acid to give an aqueous solution of H_2SiF_6 was measured by E. C. King [15]:

$$Si(\text{cryst.}) + O_2(g) + 6HF(aq.) \rightarrow H_2SiF_6(aq.) + 2H_2O(liq.) \qquad -\Delta H = 250{\cdot}3 \pm 0{\cdot}3$$

$$SiO_2(\alpha\text{-quartz}) + 6HF(aq.) \rightarrow H_2SiF_6(aq.) + 2H_2O(liq.) \qquad -\Delta H = 32{\cdot}8 \pm 0{\cdot}4$$

whence for

$$Si(\text{cryst.}) + O_2(g) \rightarrow SiO_2(\alpha\text{-quartz}) \qquad -\Delta H_f° = 217{\cdot}5 \pm 0{\cdot}5$$

The value derived is close to that obtained by Margrave, and both are significantly different from the older value.

Reaction calorimetry

An alternative approach to the study of the combustion process, which involves complete oxidation, is the study of reactions in which only a few bonds of particular interest are broken or formed. This part of thermochemistry, the measurement of heats of chemical reactions other than combustion, has come to be known as reaction calorimetry. The following examples illustrate the type of information that can be obtained from measurements of heats of hydrogenation, polymerization, and hydrolysis.

Hydrogenation. During the years 1935–1939, G. B. Kistiakowsky and his colleagues [16] measured the heats of hydrogenation of a large number of olefines in the gas phase. Since that time, the measurement of the heats of hydrogenation in solution of a wider variety of olefine, acetylenes, and aromatic compounds has become possible, by suitable choice of solvent, catalyst, and calorimeter.

The approach has usually been to compare the heats of hydrogenation of a series of similar compounds, one member of which is taken as standard. For example, the heat of hydrogenation of cyclohexene in solution, $-\Delta H = 27{\cdot}1$ kcal/mole, may be selected as a standard against which to compare the heats of hydrogenation of other cyclic olefines. The lower heat of hydrogenation of cyclopentene, $-\Delta H = 25{\cdot}7$ kcal/mole, is attributed to strain in the resulting product, cyclopentane. This is due to the repulsion forces between hydrogen atoms on adjacent carbon atoms being larger in cyclopentane than in cyclohexane, in accordance with general stereochemical considerations.

J. R. Lacher [17] has considerably widened the scope of thermochemical hydrogenation studies in the gas phase by measuring heats of replacement hydrogenation (hydrogenolysis) in saturated and unsaturated organic halides. In these reactions, halogen, other than fluorine, is eliminated as hydrogen halide. He has studied the combined hydrogenation and hydrogenolysis of vinyl chloride to ethane and hydrogen chloride, and two closely related reactions using fluorine-substituted vinyl chlorides:

$$CH_2{=}CHCl + 2H_2 \rightarrow CH_3{-}CH_3 + HCl \qquad -\Delta H = 51{\cdot}2 \text{ kcal/mole}$$

$$CF_2{=}CHCl + 2H_2 \rightarrow CHF_2{-}CH_3 + HCl \qquad -\Delta H = 61{\cdot}8 \text{ kcal/mole}$$

$$CF_2{=}CFCl + 2H_2 \rightarrow CHF_2{-}CH_2F + HCl \qquad -\Delta H = 64{\cdot}9 \text{ kcal/mole}$$

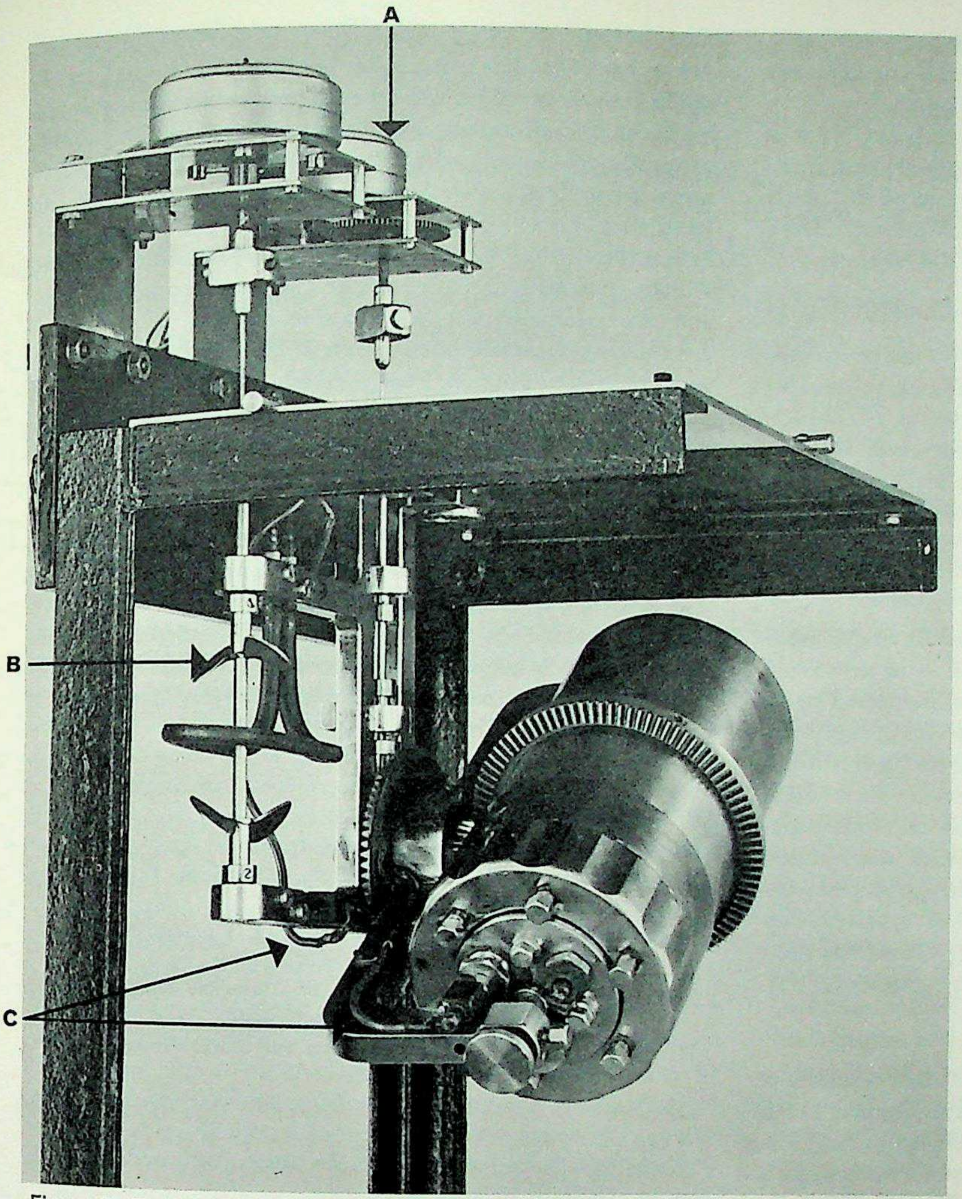


Figure 2 Rotating-bomb calorimeter with water-bath removed. A, Motor driving rotational mechanism, B, stirrer; C, wiring for electrical firing current.

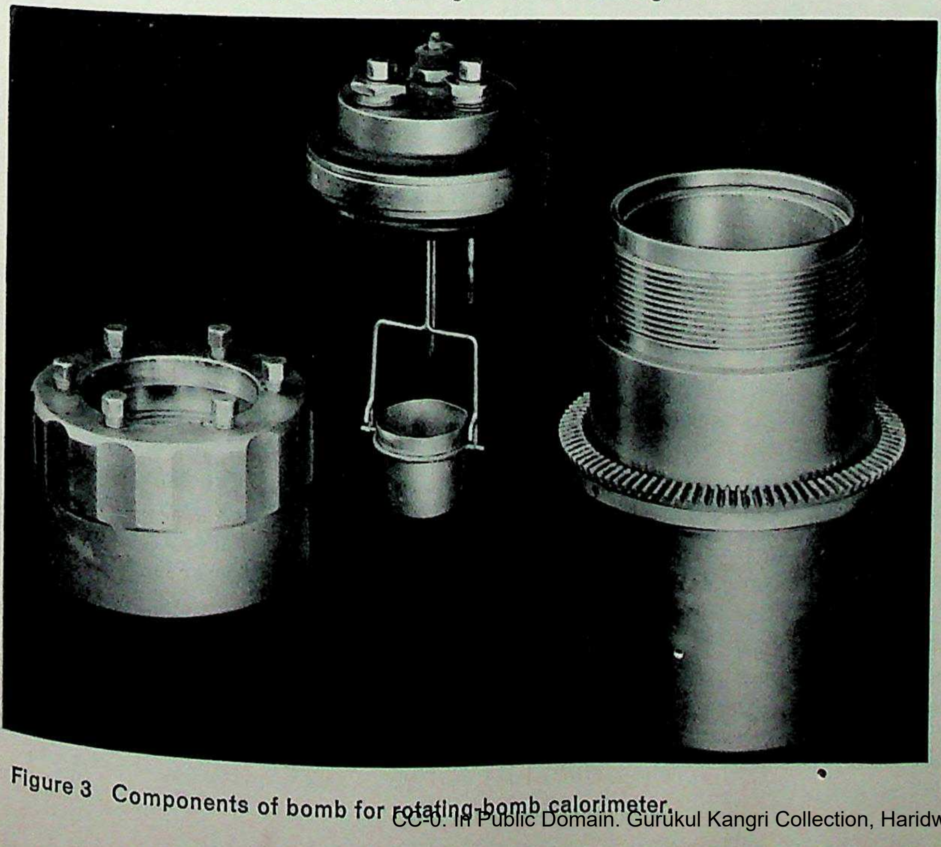


Figure 3 Components of bomb for rotating-bomb calorimeter.

The increase in the heats of these reactions with increasing fluorine content indicates increasing stabilization of the saturated reaction product relative to the unsaturated reactant.

Polymerization. Heats of polymerization of a large number of compounds have been measured by F. S. Dainton and his colleagues [17]. An example is the addition polymerization of olefine monomers of the type

$$n\text{RR}'\text{C}=\text{CR}''\text{R}''' \rightarrow (\text{—RR}'\text{C—CR}''\text{R}''')_n$$

The observed data have been correlated with molecular structure.

Hydrolysis in biological reactions. Many biological reactions necessarily occur *in vivo* in dilute aqueous solution at nearly constant temperatures in the range 25–30°C and at *p*H near to 7. They usually consist of a series of linked reversible reactions, each controlled by an enzyme catalyst, in which chemical bond energy is transferred step by step from the initiating reaction to achieve the final reaction of the series. In many such reaction chains, the penultimate stage is the release of energy by the hydrolysis of adenosine triphosphate (ATP) to adenosine diphosphate (ADP) and phosphoric acid (P). The enthalpy change of this reaction is thus a fundamental quantity in the energetics of biological reaction systems.

J. M. Sturtevant [18] and his collaborators have measured this enthalpy change in aqueous solution, using twin calorimeters within a single adiabatic shield, a method originally used by J. P. Joule in 1845. Each calorimeter *A* and *B* was charged with the same quantity of aqueous ATP. Myosin was added to *A* to start the hydrolysis, and the heat, slowly liberated as the reaction proceeded, was balanced by supplying electrical energy to *B*. The use of such a thermally balanced system effectively eliminates many errors. The energy supplied to *B* indicated that for the reaction

$$\text{ATP}^{4-}+\text{H}_2\text{O}\rightarrow\text{ADP}^{3-}+\text{HPO}_4^{2-}+\text{H}^+$$

$-\Delta H=5{\cdot}3$ kcal/mole. This corresponds closely to the value for the heat of hydrolysis of aqueous inorganic pyrophosphate ($-\Delta H=5{\cdot}8$ kcal/mole).

Earlier determinations of $-\Delta H$ for the ATP → ADP + P hydrolysis had suggested much higher values, in the range 12–15 kcal/mole, and the difference between these and that for inorganic pyrophosphate had been difficult to understand. Even the relatively low values of $-\Delta H$ now accepted for pyrophosphate hydrolysis are greater than those obtained for corresponding simple phosphate esters, such as glucose-6-phosphate, and this is the basis for the very useful concept of the high-energy phosphate bond.

Noble gas compounds. Calorimetric studies of novel types of reaction, akin to hydrolysis, have been used recently in estimating the stabilities and bond energies of several of the newly discovered noble gas compounds. Thus S. R. Gunn and S. M. Williamson [19] have measured the heat of the reaction

$$XeF_4(\text{cryst.}) + 4KI(\text{aq.}) \rightarrow Xe(\text{g}) + 4KF(\text{aq.}) + 2I_2(\text{cryst.})$$

(an overall reduction of XeF_4 to $Xe + 4F^-$ by iodine anions) and they have derived values for $-\Delta H_f^\circ$ (XeF_4, cryst.) and for $\bar{E}(Xe-F)$ of 60 and ~31 kcal/mole respectively. N. Bartlett [20], from a consideration of the thermochemical data for existing inert gas compounds, has been able to suggest which compounds of this class, apart from those already known, are likely to be stable enough to be isolated.

Microcalorimetry

Many interesting heat changes are relatively small, and in addition, the scarcity of reactants may necessitate using very small quantities. This had led to the development, pioneered by E. Calvet, of extremely sensitive microcalorimeters for the measurement of very small heat changes. The greatest problem is to eliminate the effects of variable heat exchange between the reaction cell and its surroundings. This is achieved by the Joule techniques that uses twin cells, in only one of which the reaction takes place, and measuring the difference in temperature between the two. With this apparatus, designed to operate up to moderately high temperatures (~200°C), Calvet has been able to measure reaction heats as small as 10^{-3} calorie [8, 17].

Flame calorimetry

For the study of the heats of combustion and of other reactions of substances that are either gases or very volatile liquids, bomb calorimeter techniques are either inapplicable or, at best, unsatisfactory. Such studies are best carried out in a flame calorimeter. The first calorimeter of this kind was described by Thomsen in 1886, but the development of a highly accurate model was due to Rossini, from 1931 to 1949 [8]. The flame calorimeter consists of a reaction vessel, usually of glass or quartz, immersed in a calorimeter can having a thermostatically controlled jacket. The reactant gases, sometimes in an inert carrier such as argon, flow at constant pressure and temperature into the vessel where the reaction is initiated electrically. The exhaust gases are cooled by a heat exchanger in the calorimeter can and are then swept out of the system for analysis. The heat change actually measured in the flame calorimeter is the enthalpy change, and needs only correction to 298°K and atmospheric pressure.

Rossini used his flame calorimeter for measuring the fundamentally important heat of combustion of hydrogen in oxygen. H. A. Skinner [21] has very recently used a similar flame calorimeter for a study of the heats of combustion of diethyl, ethyl vinyl, and divinyl ethers, and obtained results confirming the heats of hydrogenation of the unsaturated ethers reported by Kistiakowsky in 1938. Prosen [17] has used flame calorimeter techniques to measure directly the heats of formation of boron trichloride and titanium tetrachloride and tetrabromide by direct combination of the elements, and the heats of decomposition of di-, penta-, and deca-borane into boron and hydrogen. A fluorine flame calorimeter has recently been described [17].

Conclusion

It appears that, while combustion-bomb calorimetry is capable of use in the study of a single type of reaction, namely complete oxidation or fluorination, which can now be applied to a wide range of types of compound, reaction calorimetry is capable of use in the study of a wide range of reactions, each of which can be applied only to a fairly narrow range of types of compound. It seems probable that in the future, combustion calorimetry will be extended to more diverse types of compound, and reaction calorimetry to new types of reaction.

In the experimental arrangement of the calorimetric systems two trends are apparent. Firstly, bomb calorimetry is moving away from the classical technique, in which the bomb is immersed in a liquid (usually water), within an isothermal jacket; instead a technique of 'dry calorimetry' is evolving in which the bomb is embedded in a solid metallic calorimetric matrix, again within an isothermal jacket. Secondly, with those reaction calorimeters that use adiabatic vacuum jackets there is a tendency to move away from glass (Dewar) to metallic vacuum vessels. The most recent account of these and other developments is given in the Proceedings of the I.U.P.A.C. Symposium on Thermochemistry held in the University of Lund in July 1963 [22].

References

[1] Rossini, F. D. *Chem. Rev.*, **18**, 233, 1936.

[2a] Springall, H. D., White, T. R., and Cass, R. C. *Trans. Faraday Soc.*, **50**, 815, 1954;

[2b] Bedford, A. F., Carey, J. G., Millar, I. T., Mortimer, C. T., and Springall, H. D. *J. chem. Soc.*, 3895, 1962.

[3] Cottrell, T. L. 'The Strengths of Chemical Bonds' (second edition). Butterworths, London. 1958.

[4] Mortimer, C. T. 'Reaction Heats and Bond Strengths'. Pergamon Press, Oxford. 1962.

[5] Prosen, E. J. and Rossini, F. D. *J. Res. Nat. Bur. Stand.*, **33**, 255, 1944.

[6] Bedford, A. F., Beezer, A. E., Mortimer, C. T., and Springall, H. D. *J. chem. Soc*, 3823, 1963.

[7] Mortimer, C. T. *Pure appl. Chem.*, **2**, 71, 1961.

[8] 'Experimental Thermochemistry', Vol. I, Rossini, F. D. (Editor). Interscience Publishers, New York. 1956.

[9] Good, W. D., Fairbrother, D., and Waddington, G. *J. phys. Chem.*, **62**, 853, 1958.

[10] Good, W. D., Lacina, J. L., De Prater, B. L., and McCullough, J. P. *Ibid.*, (In press.)

[11] Wagman, D. D., Kilpatrick, J. E., Taylor, W. J., Pitzer, K. S., and Rossini, F. D. *J. Res. Nat. Bur. Stand.*, **34**, 143, 1956.

[12] Humphrey, G. L. and King, E. C. *J. Amer. chem. Soc.*, **74**, 2041, 1952.

[13] Wise, S. S., Margrave, J. L., Feder, H. M., and Hubbard, W. N. *J. phys. Chem.*, **66**, 381, 1962.

[14] Good, W. D. *Ibid.*, **66**, 380, 1962.

[15] King, E. C. *J. Amer. chem. Soc.*, **73**, 656, 1951.

[16] Conant, J. B. and Kistiakowsky, G. B. *Chem. Rev.*, **20**, 181, 1937.

[17] 'Experimental Thermochemistry', Vol. II, Skinner, H. A. (Editor). Interscience Publishers, New York. 1962.

[18] Podolsky, R. J. and Sturtevant, J. M. *J. Amer. chem. Soc.*, **73**, 2454, 1951.

[19] Gunn, S. R. and Williamson, S. M. *Science*, **140**, 117, 1963.

[20] Bartlett, N. *Endeavour*, **23**, 3, 1964.

[21] Pilcher, G., Skinner, H. A., Pell, A. S., and Pope, A. E. *Trans. Faraday Soc.*, **59**, 316, 1963.

[22] 'Symposium on Thermodynamics and Thermochemistry', Lund, July 1963. Butterworths, London. (In press.)

Weather forecasting by electronic calculation

R. C. Sutcliffe

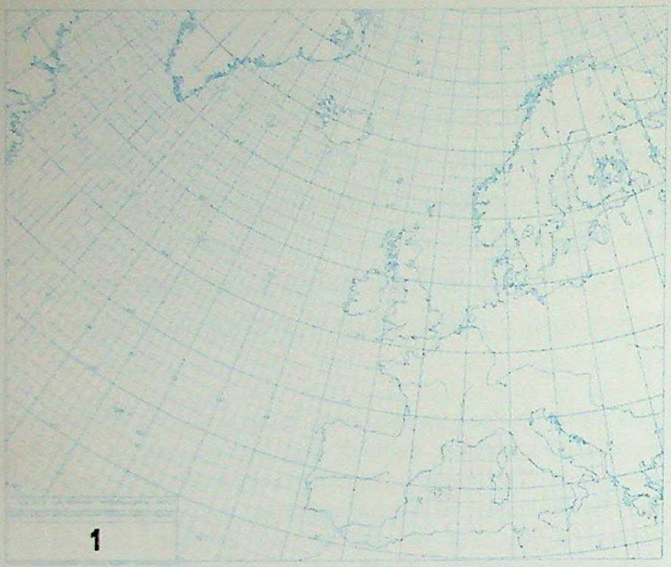
Weather forecasting has to do with the prediction of physical changes, so that one might expect the forecasts to be rigorously deducible from the conditions at any time. However, the number of calculations involved, their mathematical difficulty, and the complications involved by the 'feedback' of the changes into the system, have made the calculations effectively impossible by conventional methods. The calculations would seem, however, to be amenable to computers, and this article discusses the technique of computed forecasts and possible lines of future development.

Meteorology is, of course, in principle an exact science. It is a branch of physics—more specifically of geophysics—and it deals with quantities that can be measured, and so may be represented by numbers. Its task is to describe in these terms the behaviour of the physical system with which it is concerned, the atmosphere, and to discover the quantitative laws that govern this behaviour. We might expect that routine weather forecasting would be performed by the application of the known meteorological laws to the conditions of the day. It would then indeed be an objective science. It has, however, developed over the last hundred years quite differently, as a professional skill rather than as an objective scientific operation. The reason for the difference is not far to seek: meteorology is a branch of fluid mechanics and fluid thermodynamics, governed by known basic laws; but the boundary conditions and internal processes are of such complexity as to prohibit detailed analysis.

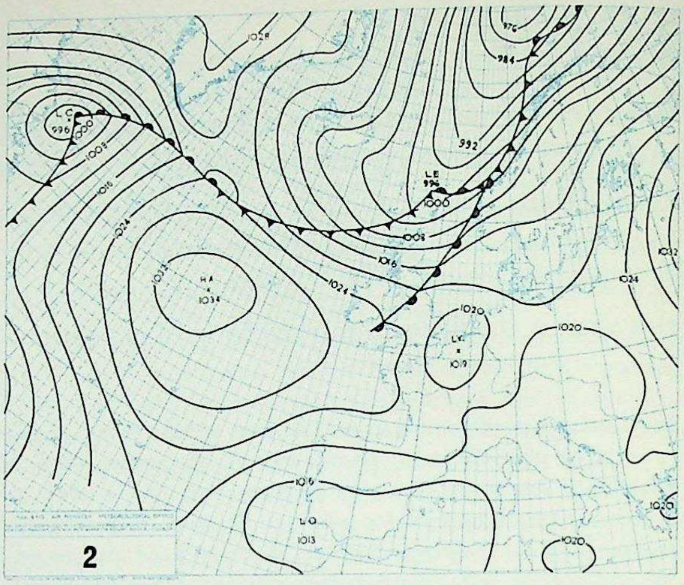
The complexity of the problems of prediction is evident enough if we run our minds over the many phenomena which go to make up the totality of a day's weather. Admittedly we are concerned with fluid mechanics, but the fluid moves on the surface of a rotating sphere to which it is held by gravity and irregularly heated and cooled. Motions are set up over a wide spectrum of scales in space and time, from the continental or oceanic wind systems to the smallest eddies, and are fully three-dimensional—or four-dimensional, if time is included. The fluid is a gas, thermodynamically active, that is, subject to heat changes, non-adiabatic changes, and also adiabatic processes that carry equal weight. The fluid does not even have a uniform composition, for there is a highly variable proportion of water vapour, as well as liquid water and ice held in suspension; and the evaporation and condensation of water contribute, through their latent heat, the largest terms in the complicated budget of energy change. Radiation from the Sun is the ultimate driving force; but its interaction with the atmosphere almost defies detailed analysis, and there are two further features of far-reaching consequence. The first of these is feedback occurring in many guises, typified by the fact that the clouds that occur in the atmosphere reflect about 40 per cent of all the Sun's heat. Thus we may hardly begin to compute energy input until we know the weather conditions. Then there is the interaction between different scales of motion, perhaps the most subtle difficulty of all.

There have, of course, always been those who regarded basic scientific methods as appropriate in principle. Indeed, probably the majority of meteorologists are deterministic in their philosophy to that extent, and in 1922 L. F. Richardson actually described a method and carried out an arduous calculation. This merely emphasized the scale of the task, not only by the failure of the one example, which was in fact not beyond repair, but by the author's estimate that 64 000 employees would be needed to keep the calculations in step with the weather changes. Instead of pioneering new methods, Richardson, in spite of his quite remarkable insight, really succeeded in laying them to rest for a generation by a sort of *reductio ad absurdum*. No one at that time foresaw electronic machines capable of computing even faster than the army of 64 000. For one who learned his meteorology and practised the art of forecasting soon after Richardson's work was published, it is fascinating to try to recover the mental attitude of the time. The greatest news of the day then was from the Norwegian school of synoptic meteorology, whose ideas on fronts and air masses were gaining general acceptance. The basic concept of this school was that the Earth's atmosphere could be divided geographically into large regions, occupied by different and more or less homogeneous air masses that were separated by boundary surfaces called fronts, continuously in motion. Weather was then divided into 'air mass weather' and 'frontal weather'. The important processes were mainly convection from the surface and up-sliding at the fronts. This scientific reasoning with its physical models brought a much-needed stimulus to forecasting, and much that was then discovered remains true for all time. However, the model of moving air masses separated by discontinuities never became amenable to dynamical treatment and progress was, in this respect, inhibited. The Norwegian ideas, after providing a great step forward, were in fact leading to a dead end, and another breakthrough was necessary to prepare the way for prediction by dynamical calculation. The breakthrough came, as it so happened, as a consequence of the wartime demands on meteorology for aviation and the need for balloon soundings of the upper air. Radio-sounding methods were devised. Attention was now directed to the winds in the higher levels of the troposphere and over large distances, and the emphasis moved away from fronts and air masses. Dynamical thinking was advancing in a number of countries, and as soon as the first electronic computer was installed at Princeton, in the United States, meteorologists tried to apply it to solving the prediction equations of dynamical meteorology. In 1946, a definite project was begun by J. von Neumann and J. G. Charney, in consultation with C. G. Rossby, and the same line of research was taken up in a number of other countries. Work with an electronic computer began in the Meteorological Office in England in 1951, under my direction. Since then, steady progress has been made, so that today we are ready to introduce the methods into regular routine use as soon as a complete computer system of adequate capacity and speed can be installed.

R. C. Sutcliffe, *C.B., O.B.E., Ph.D., F.R.S.*
Was born in 1904 and educated at Whitcliffe Mount Grammar School, Cleckheaton, Yorkshire; at Leeds University; and University College, Bangor. He entered the Meteorological Office, London, in 1927 and worked mainly with the Royal Air Force until 1946, when he was Chief Meteorological Officer in Germany. On resuming his civilian career in the Meteorological Office he engaged in full-time research and for the past ten years has been Director of Research.



1



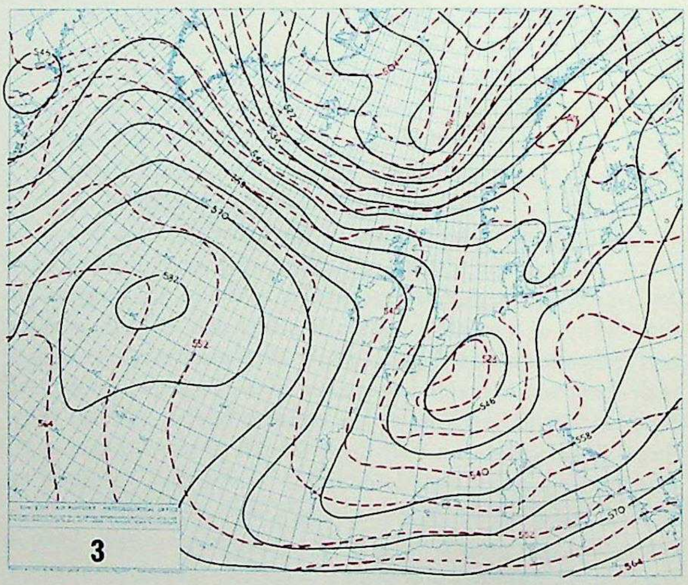
2

Figure 1 Map showing the area covered by the charts, figures 2–8.

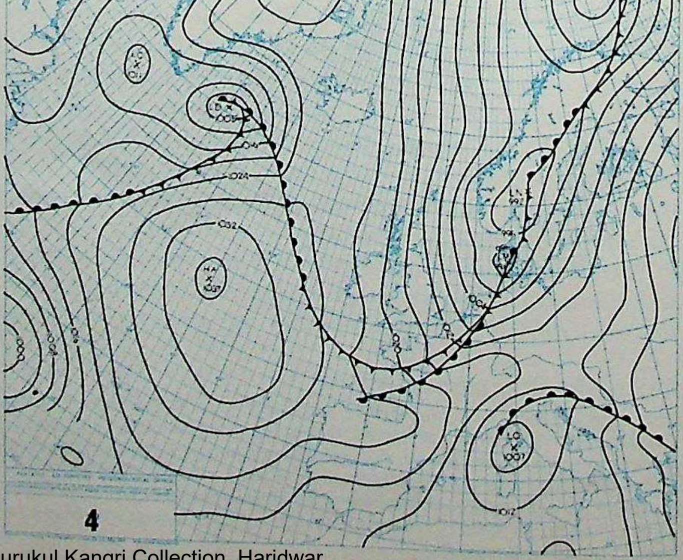
Figure 2 Surface chart for 0000 G.M.T., 14 November 1962 showing an anticyclone over the Atlantic, a depression to north of Norway with a warm front ◠◠◠ across England and a cold front ▲▲▲ extending to another depression south west of Greenland.

Figure 3 500 mb contours (black) and 1000–500 mb thickness (red) 0000 G.M.T., 14 November 1962. Numbers show heights and thicknesses in decametres. Note the cold low over France, the warm tongue extending from the Atlantic across Scotland, and the strong gradient for westerly winds to the north.

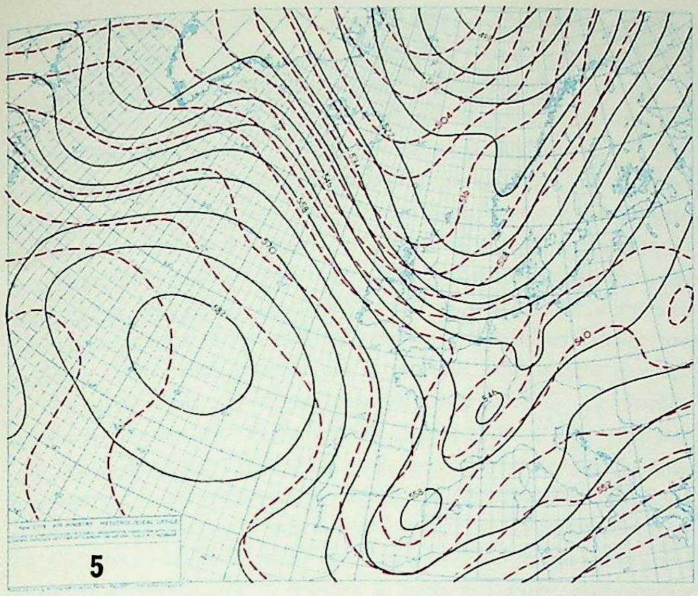
Figure 4 Surface chart for 0600 G.M.T., 15 November 1962. The anticyclone has intensified a little and the cold front has swept southwards to France, followed by northerly winds from the polar seas to the British Isles.



3



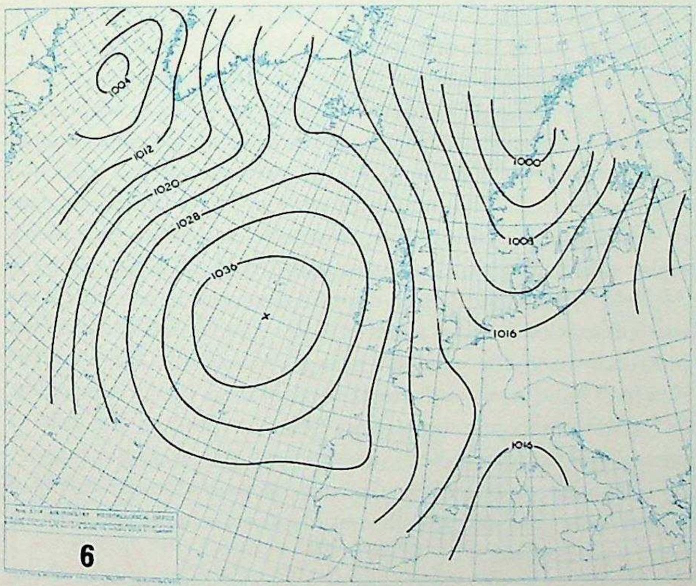
4



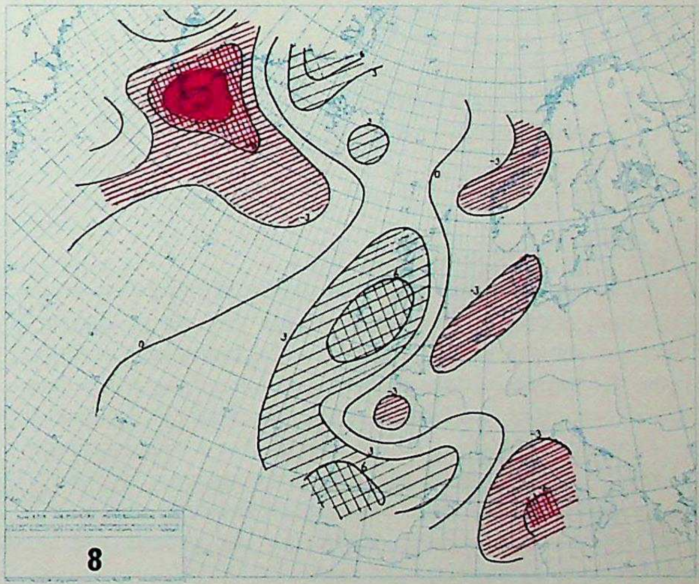
5

Figure 5 500 mb contours and 1000–500 mb thickness 0000 G.M.T., 15 November 1962. The thickness lines show clearly how the cold air has spread southwards to the British Isles, carrying the strong gradient southwards.

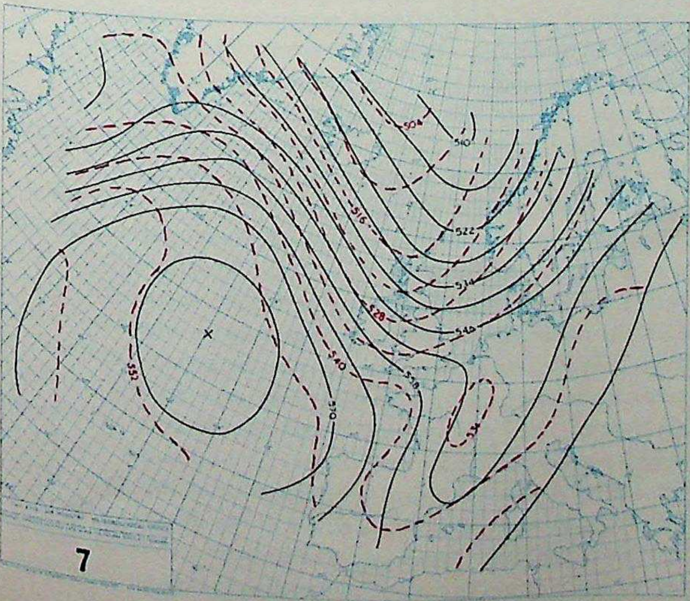
Figure 6 Computed forecast chart for 0600 G.M.T., 15 November 1962, corresponding with the actual chart of Figure 4. Pressure is not low enough over Denmark, but the essential change is well predicted.



6



8



7

Figure 7 Computed forecast chart for 0000 G.M.T., 15 November 1962, corresponding with the actual chart of figure 5. The differences are of minor importance.

Figure 8 Computed forecast of vertical velocity 0600 G.M.T., 15 November 1962. Mean for layer 1000–600 mb in mb/hr. Ascending motion (black shading) corresponds with the depression over Norway and Denmark and the trough across France. The subsidence (red shading) behind the cold outbreak over British Isles is also well indicated. Note also the predicted upward motion near Greenland and over the Mediterranean.

Dynamical meteorology

We may perhaps best begin with the basic laws of fluid dynamics and thermodynamics presented in a form likely to be familiar to physicists generally. The four variables defining the state and motion of the atmosphere are the pressure p, the temperature T, the density ρ, and the three-dimensional vector velocity $\mathbf{V}$. It is convenient to define a fifth variable, potential temperature, θ, which is the temperature obtained by compressing the air adiabatically to a standard pressure p_0, given by the equation

$$T=\theta(p/p_0)^{R/C_p} \qquad \ldots.(1)$$

where C_p = specific heat at constant pressure.

The equation of state is

$$p=RT\rho \qquad \ldots.(2)$$

The equation of continuity is

$$\frac{\partial\rho}{\partial t}=-\nabla\,.\,\rho\mathbf{V} \qquad \ldots.(3)$$

The thermodynamical equation relates change of potential temperature with the rate of heat input dq/dt:

$$\frac{\partial\theta}{\partial t}=-\mathbf{V}\,.\,\nabla\theta+\frac{\theta}{TC_p}\cdot\frac{dq}{dt} \qquad \ldots.(4)$$

Finally there is the dynamical equation referred to axes rotating with the Earth's angular velocity $\mathbf{\Omega}$:

$$\frac{\partial\mathbf{V}}{\partial t}=-\mathbf{V}\,.\,\nabla\mathbf{V}-2\mathbf{\Omega}\times\mathbf{V}-\nabla p/\rho+g+F \qquad \ldots.(5)$$

where F is the frictional term.

These five equations with five dependent variables are, in principal, a complete set, and predict the rate of change of state and motion in terms of initial and boundary conditions. From the known initial conditions, and assuming a knowledge of the heat input and friction term, equations (3), (4), and (5) give the rates of change of ρ, θ, and $\mathbf{V}$. For a sufficiently short interval they will, therefore, define new values ρ_1, θ_1, and $\mathbf{V}_1$. From equations (1) and (2) the corresponding values of T_1 and p_1 may be obtained. Repeating the process, we may proceed step by step into the future.

This simple argument from the basic equations may perhaps be taken as sufficient to indicate how prediction is possible in principle, but it does little more than this. The equations, which have been known since Euler, are not well conditioned for the description of meteorological motions. Incidentally, nothing has been introduced to specify or restrict the kinds of motion, and the equations will therefore be applicable not only to meteorological motions of all scales, from the small cumulus cloud to the general circulation of the atmosphere, but also to sound waves, pure gravity waves, and a number of other atmospheric disturbances which have no effect on weather. The problem therefore has been to adapt the equations, by suitable approximations, to the type of motion that it was hoped to be able to predict. This is the large-scale motion of the atmosphere, on the scale of depressions and anticyclones and the corresponding 'long waves' in the upper atmosphere that have a 'wavelength' or scale factor upwards of 1000 km. This scale of motion is of central importance in forecasting and is, as it so happens, the smallest scale that can be observed in reasonable detail in three dimensions by present techniques of balloon-sounding and with present financial resources.

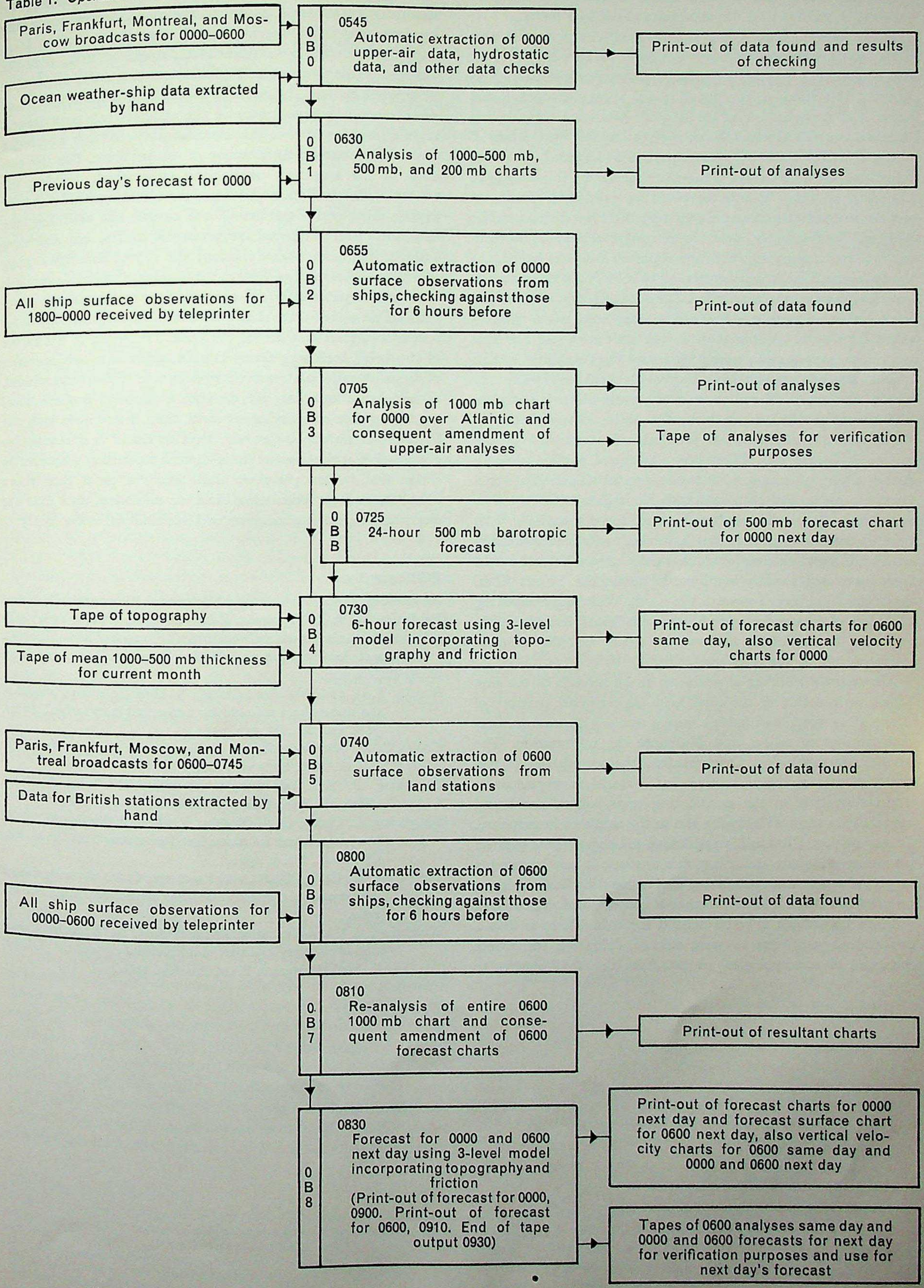
It will not be possible to mention here all the difficulties that have been encountered and more or less overcome, but two points are specially worth noting. The first is that, to a very high degree of approximation, the pressure everywhere may be treated as hydrostatic. The dynamical equation for vertical motion then becomes the hydrostatic equation $\partial p/\partial z=-g\rho$, which, by allowing ρ to be easily eliminated, effectively reduces the number of variables by one. The approximation also makes it convenient to refer position in the vertical plane to pressure rather than to linear height or, in other words, to use pressure as an independent variable and the vertical height as dependent. Upper air meteorological charts and predictions are therefore made for pressure levels of 700 mb, 500 mb, 300 mb and so on, instead of for defined horizontal surfaces (figure 3). The vertical separation from, say, 1000 mb to 500 mb, the '1000–500 mb thickness', is evidently a function only of the temperature of the layer, and 'thickness lines', the isopleths of thickness, are effectively isotherms. The mode of representation of the three-dimensional atmosphere by contour lines of pressure surfaces, and by thickness lines obtained by subtraction, thus provides a representation of both pressure and temperature, and density does not need to be introduced specifically. This system proves to be very convenient.

The second difficulty is that of handling the horizontal motion, the wind, on this scale. Difficulties of observation and the complication due to smaller-scale turbulence make it impossible to use the raw wind-data as initial conditions in the equation of continuity or in the dynamical equation. It is indeed very difficult to carry the wind as a fully independent variable in computations without introducing spurious motions and instabilities. Various ways of overcoming this difficulty have been tried. The most successful so far is the elimination of the wind in explicit form from the equations by using a scalar stream function, giving a non-divergent wind related to the field of pressure. The horizontal divergence, essential if the vertical motion is to be taken into account, can be determined indirectly from the rate of change of vorticity. This kind of sleight of hand is successful at the cost of some heavy computing. In order to recover the stream function (and so the field of pressure) from a predicted vorticity field, it is necessary to solve, by relaxation or other iterative process over the whole field, a second-order non-linear differential equation, and to do this for each time interval.

Electronic prediction in practice

The following is an outline of the method at present used in the British Meteorological Office. The geographical area covered is from Hudson Bay eastwards to Russia and is represented for computational purposes by a grid of 24 × 20 points, a mesh size of about 300 km. The structure in the vertical is represented by three pressure levels, 1000 mb, 500 (or 600) mb, and 200 mb. The parameters carried as dependent variables are a stream function ψ for the middle level (defined by a balance equation related with the height h_m of the level) and the thicknesses of the lower and upper layers, h_0' and h_1'. The two-dimensional Laplacians of these parameters are the vorticity components $\xi_m=\nabla^2\psi$; $\xi_0'=(g/f)\nabla^2 h_0'$; $\xi_1'=(g/f)\nabla^2 h_1'$ which are used in the prediction equation (f is the Coriolis parameter). The model assumes as boundary conditions that the vertical velocity vanishes at 200 mb, and that at 1000 mb it is given by a special term representing frictional divergence and topography. Parabolic profiles of vertical velocity are assumed for each layer. Non-adiabatic heating is introduced by an empirical adjustment into the lower layer, but no other account is taken either of the liberation of latent heat or of radiation exchange. The forecast is carried through for 30 hours into the future, advancing in 1-hour time steps.

Table 1. *Operational experiment in numerical forecasting.* All times in G.M.T.


Paris, Frankfurt, Montreal, and Moscow broadcasts for 0000–0600
Ocean weather-ship data extracted by hand
OB0
0545 Automatic extraction of 0000 upper-air data, hydrostatic data, and other data checks
Print-out of data found and results of checking
Previous day's forecast for 0000
OB1
0630 Analysis of 1000–500 mb, 500 mb, and 200 mb charts
Print-out of analyses
All ship surface observations for 1800–0000 received by teleprinter
OB2
0655 Automatic extraction of 0000 surface observations from ships, checking against those for 6 hours before
Print-out of data found
OB3
0705 Analysis of 1000 mb chart for 0000 over Atlantic and consequent amendment of upper-air analyses
Print-out of analyses
Tape of analyses for verification purposes
OBB
0725 24-hour 500 mb barotropic forecast
Print-out of 500 mb forecast chart for 0000 next day
Tape of topography
Tape of mean 1000–500 mb thickness for current month
OB4
0730 6-hour forecast using 3-level model incorporating topography and friction
Print-out of forecast charts for 0600 same day, also vertical velocity charts for 0000
Paris, Frankfurt, Moscow, and Montreal broadcasts for 0600–0745
Data for British stations extracted by hand
OB5
0740 Automatic extraction of 0600 surface observations from land stations
Print-out of data found
All ship surface observations for 0000–0600 received by teleprinter
OB6
0800 Automatic extraction of 0600 surface observations from ships, checking against those for 6 hours before
Print-out of data found
OB7
0810 Re-analysis of entire 0600 1000 mb chart and consequent amendment of 0600 forecast charts
Print-out of resultant charts
OB8
0830 Forecast for 0000 and 0600 next day using 3-level model incorporating topography and friction (Print-out of forecast for 0000, 0900. Print-out of forecast for 0600, 0910. End of tape output 0930)
Print-out of forecast charts for 0000 next day and forecast surface chart for 0600 next day, also vertical velocity charts for 0600 same day and 0000 and 0600 next day
Tapes of 0600 analyses same day and 0000 and 0600 forecasts for next day for verification purposes and use for next day's forecast

Note that with a more advanced computer, the time taken for this programme will be less than half an hour.

The number of elementary operations carried out in the computer is 20 million and the total time about four hours.

At the present time, a Ferranti Mercury computer is in use. This has given excellent service within its specification and capacity, but it is simply not large enough or fast enough to support this kind of computation on a regular operational basis. The introduction of an English Electric KDF9 during the next twelve months will cut the time needed by a factor of ten, and will also allow a greater area to be handled. The time schedule shown in Table 1 is therefore merely an illustration of the kind of thing that is involved in producing predicted weather maps by an almost completely objective and automatic process. The flexibility of the digital computer permits the data received by teleprinter in the form of punched tape to be fed into the machine, which is programmed to select only the relevant data and to reject the remainder. The data selection and analysis have presented problems that are in some ways more troublesome for the computer than is the mathematics. For this reason, the proving of a new programme for a modified system of analysis and computation may provide work for experts over weeks or months. Of course, once the programme is complete and checked, daily use requires only routine supervision.

An example of an actual prediction is illustrated by the charts in figures 1–8. These give the actual surface pressure patterns, 500 mb contours, and 1000–500 mb thickness patterns for the mornings of 14th and 15th November 1962, together with the numerically predicted charts for comparison. It is evident that the computations have dealt rather satisfactorily with an interesting case in which a cold current behind a cold front extended rapidly southwards across the British Isles, while the wind veered towards the north. While the computing technique does not deal specifically with fronts, their positions are roughly implied by the pressure fields, and the air mass temperatures by the thickness fields.

There are a number of problems to be solved. In the first place, we must cover the earth with an adequate number of observation posts, and must arrange world-wide communication systems to collect and distribute the information. The World Meteorological Organization, the association of virtually all meteorological services in the world, is very seriously addressing itself to the outstanding gaps, which are found especially in tropical latitudes and in the southern hemisphere. It may be that £100 million will need to be spent each year on obtaining adequate data, but this amount is not impossibly large. Adequate data processing and computing facilities must be provided in a sufficient number of centres, and some tens of millions of pounds may be needed for this. There is also a tremendous challenge to pure science. Through the whole spectrum of meteorological events, from the small shower to the weather of the season or the climate of the century, the mathematical physicist has solved the prediction problem only by exploiting what one might call a single soft spot. It so happens, and this by a tremendous stroke of luck, that the changes in the pressure and wind patterns on the large scale of depressions and anticyclones take place, for a few days at most, almost independently of the smaller-scale and larger-scale changes, and it is this circumstance that has allowed a reasonably accurate formulation of the problem. But the two ends of the spectrum are almost untouched. So far, for example, the prediction of the pressure field, valuable as it is, says nothing about the fronts, the clouds, the visibility, the maximum and minimum temperatures, and so on, and very little, if anything, about rainfall: the expert forecaster must still add all these things. And at the other end, the new methods have made practically no contribution to long-range forecasting, an art of little skill that has been languishing under scientific neglect all through the years. However, in a number of countries, including Great Britain, there is a new measure of hopefulness, and research groups are being established. Although no one has yet devised a physical-mathematical model capable of carrying us more than a few days into the future, and no one is yet sure that to do so is a reasonable expectation at the present time, there is no reason whatever to doubt that in ten years we shall know a great deal more about long-range forecasting than we do today, and that the electronic computing machine will have helped enormously.

Bibliography

An excellent monograph entitled 'Numerical Weather Analysis and Prediction' by P. D. Thompson (Macmillan, New York, 1961) contains most of the important references up to the time of its publication.

Additional research papers referring to more recent work in Great Britain are:

Bushby, F. H. and Whitelam, Clare J. 'A three-parameter model of the atmosphere suitable for numerical integration'. *Quart. J. R. met. Soc.*, 87, 373, 1961.

Corby, G. A. 'Some experiments in the objective analysis of contour charts'. *Ibid.*, 87, 371, 1961.

Graystone, P. 'The introduction of topographic and frictional effects in a baroclinic model'. *Ibid.*, 88, 377, 1962.

Knighting, E., Corby, G. A., Bushby, F. H., and Wallington, C. E. 'An experiment in numerical forecasting'. *Sci. Pap. met. Off., Lond.*, No. 5, 1961.

Knighting, E., Corby, G. A., and Rowntree, P. R. 'An experiment in operational numerical weather prediction'. *Ibid.*, No. 16, 1962.

Knighting, E. 'Numerical weather prediction'. In 'Numerical solution of ordinary and partial differential equations' (Editor: L. Fox). Pergamon Press, London. 1962.

Teichoic acids and the bacterial cell wall

J. Baddiley

The walls of bacteria are associated with their sensitivity to the penicillins, and with serological reactions. Study of the walls has led to the discovery of a new group of natural polymers, the teichoic acids, that are present in the wall and immediately inside it. This article describes the separation of the teichoic acids, and discusses their significance as a source of antigenic properties.

When bacteria are examined under a microscope or an electron microscope they are seen to possess characteristic physical features [1]. These include their overall shape and the presence or absence of flagellae, hair-like structures, gelatinous capsules, and internal granules. Such features are recognized in the classification of the organisms. Nevertheless, although wide variations occur in the possession of these detailed features, all bacteria conform to a general structural pattern; they consist of cytoplasm enclosed by a thin fragile membrane, usually called the protoplast membrane, and a much thicker structure, the wall, surrounding this. In some cases the wall is itself surrounded by a gelatinous mass, the capsule.

The wall and its immediate vicinity are particularly important parts of the organism. The wall itself is interesting, because it must possess structural rigidity for its function in protecting the cytoplasm and its membrane from the considerable osmotic forces that bacteria encounter, yet it must be permeable to nutrients and metabolites. As it represents up to 20 per cent of the dry weight of the cell, an appreciable proportion of the metabolic activity of bacteria is concerned with the synthesis of the wall and its precursors. Its importance in bacterial metabolism has been recognized since the discovery that the main action of penicillins on bacteria is to interfere with the biosynthesis of the wall and to cause the accumulation of precursors. Another important aspect of walls relates to the behaviour of bacteria in the tissues of an infected animal; antigens in the organisms stimulate the production of antibodies in the serum, and the subsequent antigen-antibody reactions are significant in assisting the animal to overcome the infection. Many antigens are in, or associated with, the wall. It is clear therefore, that knowledge of the chemical structure of the wall is important for a good understanding of many aspects of the biology and chemistry of bacteria.

Structure and composition of cell walls

Our knowledge of the chemistry of bacterial walls has advanced rapidly in recent years, mainly since the development of satisfactory methods for separating them in reasonable amounts in the laboratory. Although these vary according to the organism under investigation, most involve mechanical rupture followed by exhaustive washing to remove all traces of cell contents. In a convenient procedure the suspension of organisms is shaken at a high speed with very small glass beads [2]. Typical wall preparations are shown in figures 2 – 6, where it is seen that the clean walls are flattened structures that retain the general shape of the original organisms. Although such preparations are now accessible, it is difficult to obtain them in large amounts, and chemical work on the polymers of which they are composed has been carried out mainly on a micro scale. Paper chromatography and enzymic methods have been extensively used for the identification of degradation products.

The composition of walls differs markedly with bacterial staining properties. Gram-negative[1] organisms possess walls that are chemically highly complex. They are known to contain much lipid, associated with protein and polysaccharide, but relatively little is known about the detailed structure of most of these complexes. On the other hand, Gram-positive bacteria mainly possess simpler walls that contain only traces of lipid and protein: the principal components are mucopeptides, soluble polysaccharides, and the more recently discovered teichoic acids.

Probably all bacterial walls contain mucopeptides, and in some Gram-positive organisms these polymers represent most of the wall material. They are highly insoluble and are responsible for structural rigidity; for example, the walls of *Bacillus polymyxa* in figures 3 and 4, which show marked detail of fine structure, are almost entirely mucopeptide. These compounds are composed of a polysaccharide chain containing acetylglucosamine and acetylmuramic acid, to which are attached short peptide side-chains as shown in figure 1 [3].

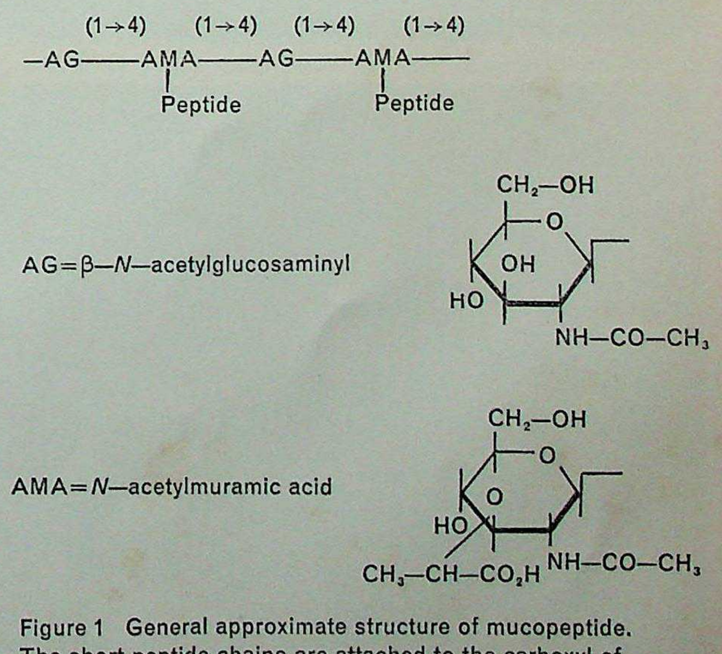
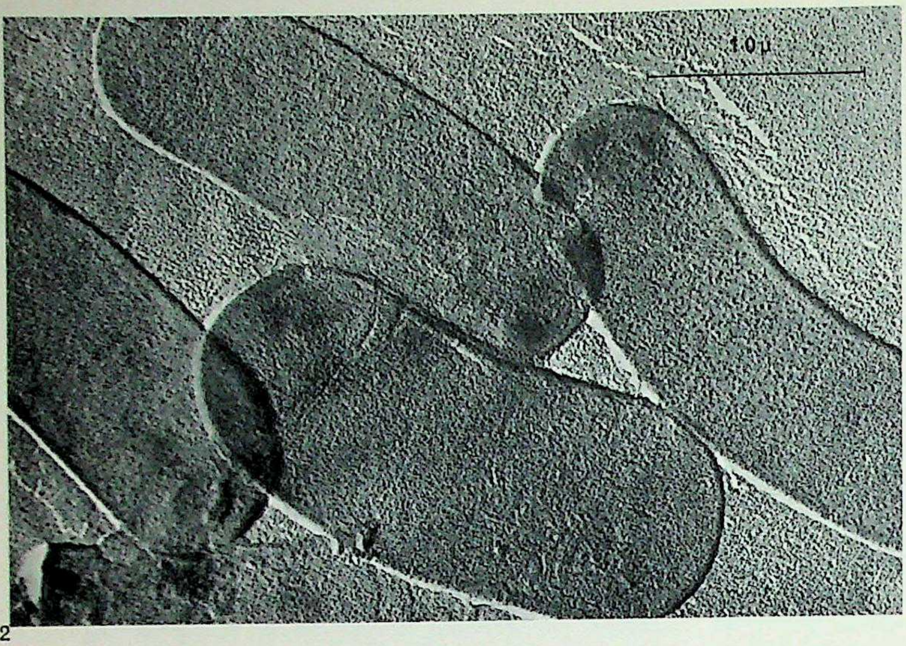


Figure 1 General approximate structure of mucopeptide. The short peptide chains are attached to the carboxyl of muramic acid; they are usually composed of alanine, glutamic acid, aspartic acid, lysine, diaminopimelic acid, and occasionally glycine and serine.

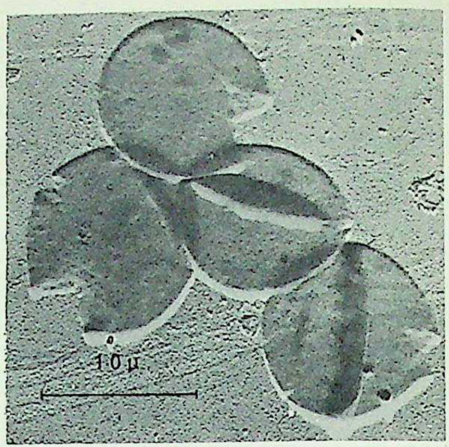
J. Baddiley, *Ph.D., D.Sc., F.R.S.*
Was born in 1918 and educated at Manchester Grammar School and Manchester University. After research at the University of Cambridge and at the Wenner-Gren Institute, Stockholm, he became a member of the staff of the Lister Institute of Preventive Medicine, in London. In 1955, he was appointed Professor of Organic Chemistry at King's College, University of Durham, now the University of Newcastle upon Tyne.

---

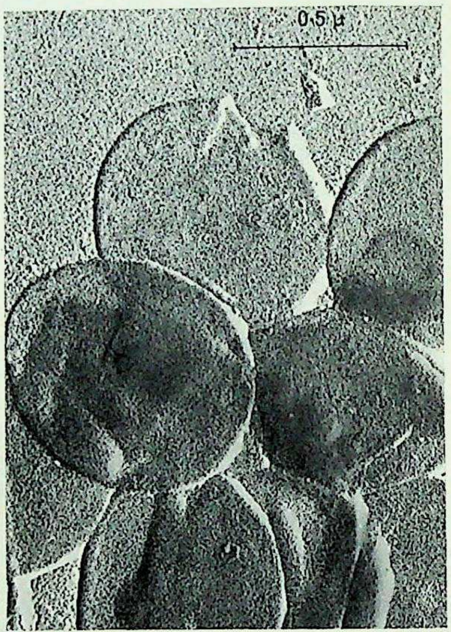
[1] In 1884, the Danish bacteriologist C. Gram introduced a staining technique that divides bacteria into two large groups. Briefly, the negative group consists of bacteria that are readily decolorized by alcohol after treating first with certain aniline dyes and then with a solution of iodine in potassium iodide; the other group is not decolorized. Although the test is an empirical one, it is nevertheless correlated with certain other characters, and the distinction is still an important one.



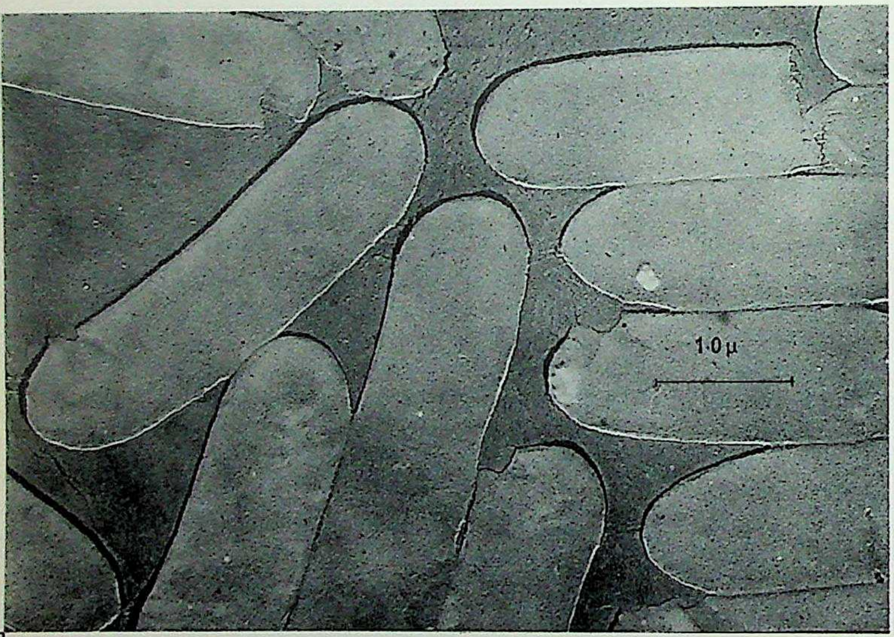
2



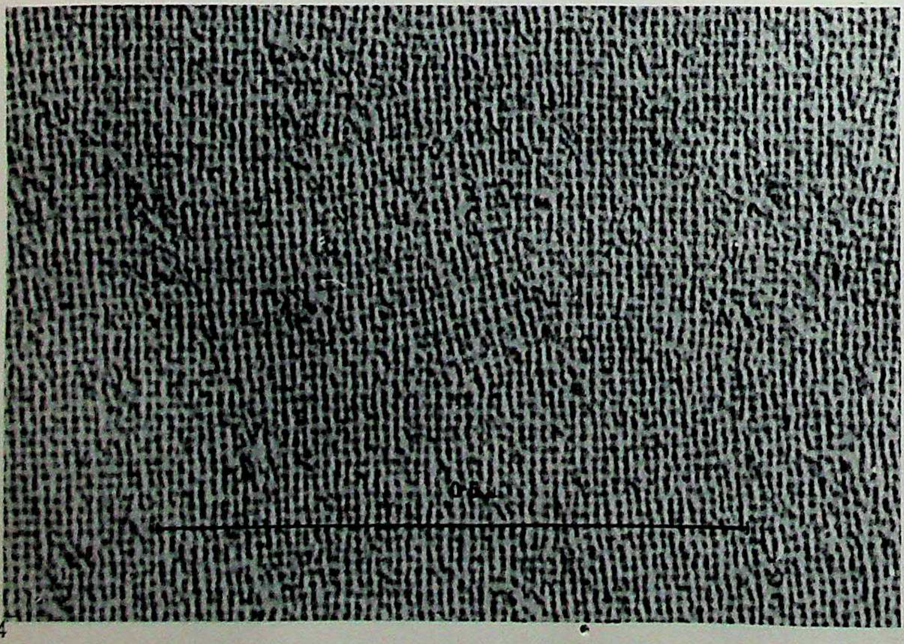
5



6



3



4

Figure 2 *Lactobacillus lactis* 7278 walls containing a glycerol teichoic acid with glucosyl substituents.

Figure 3 *Bacillus polymyxa* 4747 walls composed almost entirely of mucopeptide.

Figure 4 *Bacillus polymyxa* 4747 walls at a higher magnification. The fine structure in this wall is not typical and suggests an orderly arrangement of mucopeptide molecules.

Figure 5 *Staphylococcus epidermidis* 14 walls containing a glycerol teichoic acid with acetylgalactosaminyl substituents. The thickened band in some of the walls represents the point of cell division.

Figure 6 *Staphylococcus aureus* H walls containing a ribitol teichoic acid with acetylglucosaminyl substituents.

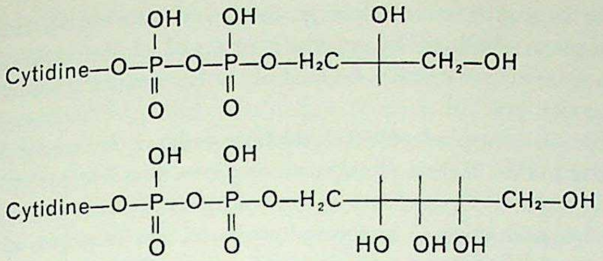


Figure 7 Nucleotides: cytidine diphosphate glycerol and cytidine diphosphate ribitol, precursors of teichoic acids.

Some of the components of mucopeptides are highly characteristic. Muramic acid is an aminosugar acid found only in bacterial walls and their precursors; diaminopimelic acid, D-alanine and D-glutamic acid are found characteristically in mucopeptides. As mucopeptides are insoluble, it is not yet possible to assess their homogeneity, and the formula in figure 1 is no more than an approximation. In fact, it is clear that mucopeptides from different organisms possess notably different general structures, and there is evidence in some cases for cross-linkages between polymer chains and between neighbouring peptide side-chains. An important feature of mucopeptides is that they are affected by penicillins; their biosynthesis is inhibited and nucleotide precursors accumulate [13].

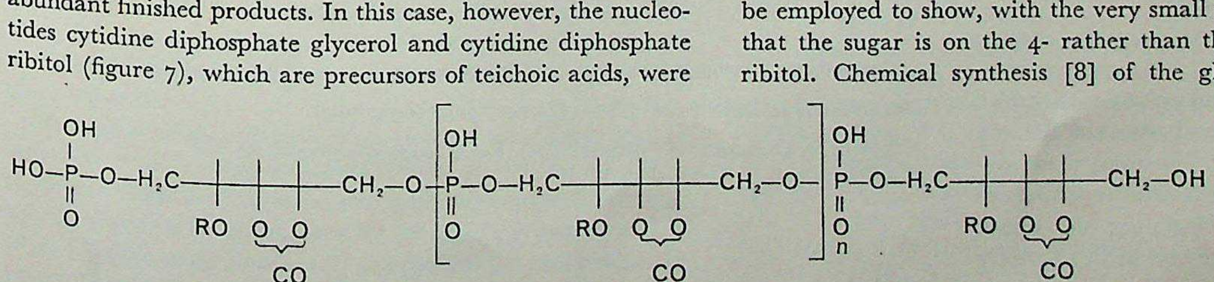
Polysaccharides have often been found in bacterial walls and capsules. After extraction, they are usually freely soluble in water, but the exact nature of the forces that hold them in the wall is not known. Their structures are sometimes complex, and some contain unusual sugars. It is not possible to discuss here the details of the chemistry of this group of compounds, many of which are serologically reactive [4].

Teichoic acids

Several years ago a new group of polymers, the teichoic acids, was discovered in the walls of a number of Gram-positive bacteria [5]. Their structure differs from that of other naturally occurring polymers in that they contain chains of glycerol or ribitol residues joined to each other by phosphate groups; sugars are attached to the glycerol or ribitol, and D-alanine residues are in ester linkage with some of the hydroxyl groups. Historically, the teichoic acids are probably unique, since their discovery was a logical consequence of the earlier discovery of two of their biological precursors. The recognition of natural precursors, which are often found only in very small amounts in living cells, generally follows the discovery of the much more abundant finished products. In this case, however, the nucleotides cytidine diphosphate glycerol and cytidine diphosphate ribitol (figure 7), which are precursors of teichoic acids, were isolated in trace amounts from bacteria, their structures determined, and their chemical synthesis completed, all before the existence of the polymers was suspected [6]. By analogy with the function of certain other nucleotides in the biosynthesis of polysaccharides, it was thought that these cytidine compounds might be required for the synthesis of polymers containing glycerol phosphate and ribitol phosphate residues. This reasoning led to the discovery of considerable amounts (up to 50 per cent of the dry weight) of such polymers in the walls of many Gram-positive bacteria, and smaller amounts of related polymers in the 'cell contents' fraction of disrupted organisms.

The teichoic acids are water-soluble compounds that can be extracted from the walls or the cells by dilute acids. The procedures that have been most valuable in structural work are hydrolysis with acid, and a combination of hydrolysis with alkali followed by dephosphorylation of products with a phosphatase [5]. Thus, hydrolysis of ribitol teichoic acids with acid yields a complex mixture of products, including alanine, inorganic phosphate, ribitol, its mono- and di-phosphates and their degradation products, together with a sugar that may vary according to the organism. A characteristic product that is readily detected on paper chromatograms of these hydrolysates is 1,4-anhydroribitol, which arises through decomposition of ribitol phosphates. The presence of anhydroribitol phosphate in acid hydrolysates, and derivatives of ribitol diphosphate in alkali hydrolysates, is consistent with a general structure of the polymers in which ribitol units are joined through phosphodiester linkages (figure 8). Hydrolysis with alkali removes the alanine, which is attached through very labile ester linkages, and then breaks the chain at each phosphodiester linkage. In this respect, teichoic acids resemble ribonucleic acids, where alkali hydrolyses the phosphate linkages through an intermediate formation of cyclic phosphates that subsequently open to give mixtures of isomeric monophosphates. Phosphate groups may be removed from the mixture of products by an enzyme. This gives high yields of ribitol glycosides, some of which have been crystallized and synthesized.

These methods can be illustrated with the ribitol teichoic acid from the walls of *Bacillus subtilis* (figure 8) [7]. In this case, the sugar is glucose, and hydrolysis with alkali and the phosphatase gives a crystalline ribitol glucoside. Oxidation of the glucoside with periodate indicates that the sugar is attached to the 4-position of D-ribitol, and enzymic and other evidence showed that the glycosidic linkage has the β-configuration. More subtle experiments and stereochemical reasoning have to be employed to show, with the very small amounts available, that the sugar is on the 4- rather than the 2-position in D-ribitol. Chemical synthesis [8] of the glucoside confirmed



Bacillus subtilis (R=β-glucosyl; n=7)

Staphylococcus aureus H (R=α- and β-*N*-acetylglucosaminyl; n=6)

Lactobacillus arabinosus 17–5 (R=α-glucosyl, and alternate ribitol residues also have α-glucosyl at the 3-position; n=4–5)

Figure 8 Ribitol teichoic acids from walls.

all the structural features that had been established by degradation. The position of the phosphate groups in the polymer followed from the separation and characterization of the isomeric monophosphates obtained by hydrolysis with alkali, together with studies of the action of periodate on the teichoic acid after removal of alanine.

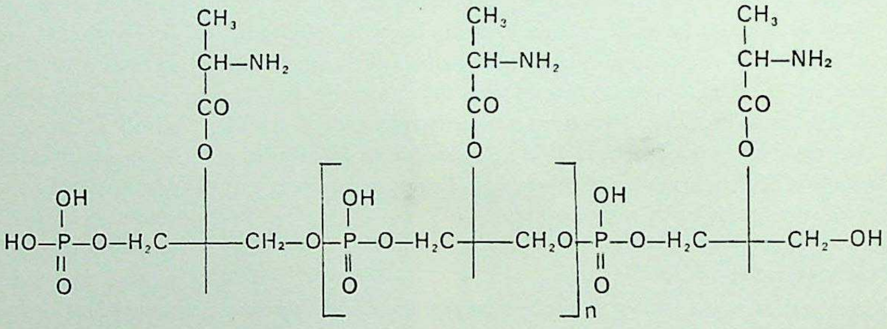
The teichoic acid from walls of *B. subtilis,* like all others examined in detail so far, can be regarded as a polymer formed by the progressive transfer of D-ribitol 5-phosphate units from the nucleotide CDP-ribitol to the 1-position of the previous unit in the chain; recent work on cell-free enzyme systems supports this view, but it is not yet clear at what stage in the biosynthesis the sugar and alanine residues are attached. The polymer chain is terminated at one end by a phosphate group and at the other by a hydroxyl group. This has been demonstrated by removal of alanine followed by three independent quantitative procedures: titration, specific hydrolysis of terminal phosphate with an enzyme, and the production of formaldehyde by oxidation of the terminal glycol with periodate. These measurements indicate that the molecule consists of nine glucosylribitol phosphate units, and other evidence suggests that the preparations are fairly homogeneous with respect to chain length. It is not possible, at present, to assign an exact position to the alanine ester residues, but chemical methods have been used to show that they must be attached to hydroxyl at the 2- or 3-positions on ribitol.

The structures of other ribitol teichoic acids have been established by methods comparable to those already described. One, which represents an appreciable proportion of the walls of *Staphylococcus aureus H* (figure 8), is similar to that from *B. subtilis.* The sugar is *N*-acetylglucosamine, mainly in β-linkage to ribitol, but a small and variable proportion of α-linkage usually occurs and the polymer contains eight ribitol units [9]. That from the walls of *Lactobacillus arabinosus* contains six or seven units bearing either one or two α-glucose substituents at positions 4, or at 3 and 4; it is likely in this case that the mono- and di-glucosyl substituents alternate regularly [9].

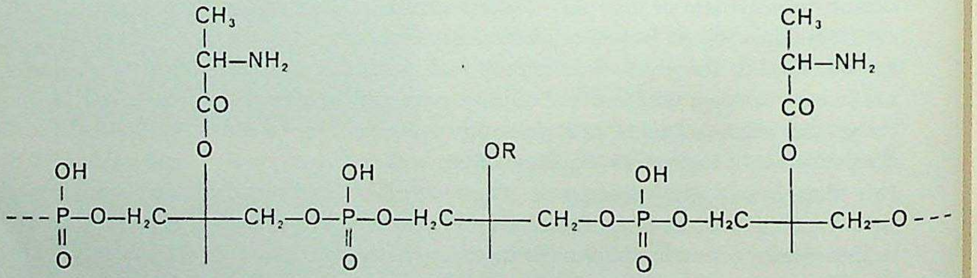
Glycerol teichoic acids

The walls of some bacteria contain teichoic acids in which the chain is composed of glycerol phosphate rather than ribitol phosphate and are thus derived from CDP-glycerol. However, these glycerol teichoic acids are much more widely distributed than are the ribitol compounds, and they occur in relatively small amounts in the 'cell contents' fraction of nearly all Gram positive organisms, regardless of the presence or absence of glycerol or ribitol teichoic acids in the walls. Such representatives have been called 'intracellular' teichoic acids, but it is now believed that they are located in the region between the wall and the underlying protoplast membrane. This interesting observation was made in studies on whole organisms, on their walls, and on protoplasts obtained by enzymic removal of the walls from organisms under conditions that allow the resulting structures to survive [9].

The chemistry of glycerol teichoic acids is in many ways similar to that of their ribitol counterparts. In a few cases, little or no sugar is present, and hydrolysis with alkali gives mainly alanine, and glycerol and its phosphates. An example of this type has been isolated from the cell-contents fraction of *L. casei* (figure 9). An interesting feature of the chemistry of this group of polymers is that, during hydrolysis with alkali, the necessary cyclization of phosphate with adjacent hydroxyl groups is not unidirectional, as it is with the ribitol compounds. This permits the formation of small quantities of diglycerol phosphate derivatives that possess no free hydroxyl groups adjacent to the phosphodiester group, and so are stable in alkali. The formation of such products can be explained only if the phosphodiester groups in the original polymer are at 1,3- and not at the 1,2- or 2,3-positions on the glycerol. Another important reaction of these polymers is their degradation on treatment successively with periodate, an organic base, and a phosphatase. In this way, it is possible to remove glycerol phosphate units in a stepwise manner from one or both ends of the polymer chain.



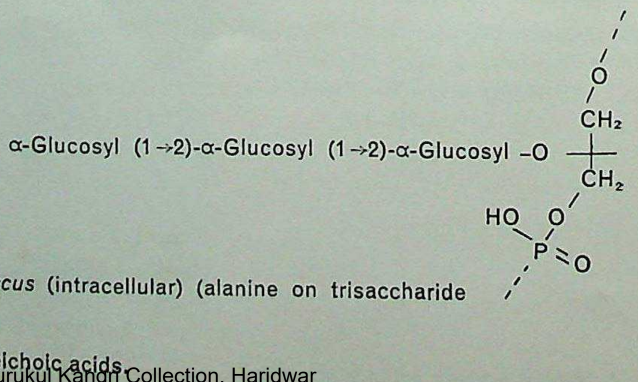
Lactobacillus casei (intracellular)



Lactobacillus arabinosus (intracellular (R=α-glucosyl)

Staphylococcus lactis (wall) (R=α-*N*-acetylgalactosaminyl)

Staphylococcus aureus H (intracellular) (R=*N*-acetylglucosaminyl and β-glucosyl [1-6]-β-glucosyl)



Group D *Streptococcus* (intracellular) (alanine on trisaccharide residue)

Figure 9 Glycerol teichoic acids

Considerable variations occur in the amount and nature of the sugar residues in glycerol teichoic acids. Those from some lactobacilli contain little or no sugar, whereas others contain oligosaccharide residues attached to each glycerol. The cell contents fraction of *L. arabinosus* contains one comprising about eighteen glycerol phosphate units, only two of which have α-glucosyl substituents (figure 9). One obtained similarly from *Staph. aureus H* contains only a few acetylglucosaminyl and gentiobiosyl residues. On the other hand, serologically important teichoic acids from group *D* streptococci contain di- or trisaccharide residues on each glycerol [10]. An intermediate type occurs as an important component of the walls of a variety of *Staph. lactis*; in this case the polymer has about eighteen glycerol phosphate residues, one-third of which possess α-*N*-acetylgalactosaminyl substituents equally spaced along the polymer chain.

General surveys show that many other teichoic acids occur in a wide variety of bacteria and that, although considerable variation may be expected with respect to sugar substituents, all contain either glycerol or ribitol. No other polyols have been found in these compounds. Nevertheless, good evidence exists for the occurrence of teichoic acids in which the sugar residues are components of the polymer chain rather than branch substituents, as in the cases described so far. In this way, they resemble some of the capsule compounds that occur in certain types of pneumococcus.

Serological reactions

It is known that the outer regions of many bacteria contain macromolecular compounds, usually proteins or polysaccharides, that are responsible for the more important serological reactions of the organism. These are the antigens that stimulate the formation in animals of highly specific antibodies. The serum from such immunized animals will form a precipitate with the antigen *in vitro*. Clearly, teichoic acids in or beneath the walls might be expected to possess such serological properties. Support for this view was first obtained during a general study of the distribution of teichoic acids in lactobacilli [11]. It was noted that, although the presence and chemical nature of teichoic acids could not be related to the general biochemical properties of the different organisms, they are related to the serological groups which comprise the lactobacilli. This work has been extended recently with M. E. Sharpe and A. L. Davison, and it is now known that the group-specific compound for several of the serological groups of lactobacilli is the teichoic acid in the wall. In other groups, the specific compound is the intracellular teichoic acid. In a few cases, the teichoic acid is not responsible for group reactions, and it is probable that specificity is then associated with wall polysaccharide.

Serological properties have been ascribed to teichoic acids in other bacteria. Thus, group-specificity in strains of *Staph. aureus* and *Staph. lactis* is associated with the wall teichoic acid, but the problem is complicated by the different proportions of α- and β-glycosidic linkages in the different *aureus* strains. In group *D streptococci* it has been shown that the group-specific compound is the intracellular teichoic acid [12].

Although it is now recognized that these compounds play an important part in the reactions that occur between many bacteria and their animal hosts, much still remains to be done. It is not yet known, for example, whether isolated teichoic acids are true antigens, that is whether they can themselves stimulate the animal to produce specific antibodies, or whether they must be associated with walls or whole organisms before they do this. Similarly, the significance of these compounds in the cause and prevention of disease is not known. More work is required on the relationship between their structure and serological specificity, although it seems that the nature and configuration of the sugar residues are probably the most important factors. It has already been found that knowledge of the chemical composition of wall teichoic acids can assist greatly in the classification of organisms within a species.

Possible function

Although considerable progress has been made in the study of the chemical structure and serological properties of this interesting group of compounds, nothing is yet known about their function. Suggestions that they may be concerned in regulating the transport of metabolites and ions, or that they may participate in the biosynthesis of mucopeptides, are still speculative, and experimental evidence is lacking. Nevertheless, the densely charged environment created in or beneath a wall by their presence must surely influence the passage of electrolytes. It might be thought that the absence of teichoic acids from the walls of a number of Gram-positive bacteria, and perhaps from most Gram-negative bacteria, would indicate that they are not important for the normal activities of organisms. However, intracellular glycerol teichoic acids have been found in very nearly all the Gram-positive bacteria examined, and their presumed absence from the Gram-negative bacteria may well be offset by the presence in those organisms of related aminoacyl phospholipids. It seems likely then that teichoic acids have a biochemical importance to the bacteria in addition to their chemical and serological importance.

References

[1] Luria, S. E. In 'The Bacteria'. Gunsalus, I. C. and Stanier, R. Y. (Editors). Vol. 1, Academic Press, New York and London. 1960;
Murray, R. G. E. *Ibid.*, p. 35;
Salton, M. R. J. *Ibid.*, p. 97.
[2] Salton, M. R. J. and Horne, R. W. *Biochim. biophys. Acta*, **7**, 177, 1951.
[3] Salton, M. R. J. and Ghuysen, J. M. *Ibid.*, **36**, 552, 1959; *Idem. Ibid.*, **45**, 355, 1960.
[4] Stacey, M. and Barker, S. A. 'Polysaccharides of Micro-organisms'. Clarendon Press, Oxford. 1960.
[5] Armstrong, J. J., Baddiley, J., Buchanan, J. G., Carss, B., and Greenberg, G. R. *J. chem. Soc.*, 4344, 1958.
[6] Baddiley, J. Tilden Lecture, *Proc. chem. Soc., Lond.*, 177, 1959.
[7] Armstrong, J. J., Baddiley, J., and Buchanan, J. G. *Biochem. J.*, **76**, 610, 1960; **80**, 254, 1961.
[8] Baddiley, J., Buchanan, J. G., and Hardy, F. E. *J. chem. Soc.*, 2180, 1961.
[9] Hay, J. B., Wicken, A. J., and Baddiley, J. *Biochim. biophys. Acta*, **71**, 188, 1963.
[10] Wicken, A. J. and Baddiley, J. *Biochem. J.*, **87**, 54, 1963.
[11] Baddiley, J. and Davison, A. L. *J. gen. Microbiol.*, **24**, 295, 1961.
[12] Wicken, A. J., Elliott, S. D., and Baddiley, J. *Ibid.*, **31**, 231, 1963.
[13] Abraham, E. P. *Endeavour*, **18**, 212, 1959.

Feeding habits of tsetse flies

B. Weitz

It has long been known that the tsetse fly is the vector for the transmission of the diseases sleeping sickness to man and nagana to domestic animals. The organisms that cause the disease, trypanosomes, live in wild animals and are collected by the blood-sucking flies when they feed. Thus both the fly and the trypanosome reservoir can be destroyed by killing off the appropriate wild animal. This article describes the techniques that have been devised for studying the individual feeding habits of tsetse flies in their natural state, an essential preliminary to this method of control.

Introduction

The close association of the tsetse fly *Glossina* with a fatal disease of cattle known as nagana was first established by Sir David Bruce in 1895, who called the condition the 'tsetse fly disease' of cattle. This description correctly emphasized the part played by the fly as the vector of the trypanosomes that infected the herds of Zululand at that time. Bruce also noted the relationship of the fly with wild animals, and his field observations indicated the importance of certain game animals as the reservoir of the parasitic trypanosomes. Thus, at a very early stage in the study of trypanosomiasis, wild animals were implicated in the disease, and the tsetse fly was incriminated as the link between the reservoir hosts and the domestic communities.

Nearly 4½ million square miles of the African continent are infested with tsetse flies, and wild life remains undisturbed in most of these vast unclaimed areas of bushland. The fly is largely responsible for the retarded development of these areas, as it is the main vector of the trypanosomes that cause disease both to man and to his domestic animals. The economy of most African countries is closely associated with local agriculture, and particularly the raising of cattle, and thus the prosperity of African populations depends on the availability of suitable grazing lands. Unfortunately, a large proportion of otherwise suitable grazing areas is made unsafe to cattle by the tsetse fly.

Control of the tsetse fly requires an accurate knowledge of the feeding habits of the insect. Once it had been established that the tsetse fly depended mainly on game animals as a source of food, it became clear that one method of destroying the fly was to starve it by killing off these animals. This method was very successful in areas where it was applied, and it had the advantage of removing both the vector and the reservoir of the trypanosomes simultaneously. However, the wholesale destruction of game, apart from ethical considerations, is an ambitious and expensive project, and other means of controlling the insects by the alteration of their habitat were therefore evolved. Successful control by any means depends on a thorough knowledge of the bionomics of the insect. For this reason, a very extensive study of the natural history of the tsetse fly was made by the pioneers in this field, and a most comprehensive documentation of its behaviour was obtained [1]. However, the feeding habits were not studied until comparatively recently, largely because of the lack of suitable techniques.

B. Weitz, *D.Sc., M.R.C.V.S.*

Was born in London in 1919 and was educated at the Royal Veterinary College, London. He started research on the bacteriology of animal corynebacteria at the Agricultural Research Council's Field Station at Compton and then at the Veterinary Laboratory of the Ministry of Agriculture, Weybridge. He joined the Lister Institute of Preventive Medicine in 1947 and became Head of the Serum Department in 1952. His main research interest is the immunology of trypanosomes.

The physiology of tsetse fly feeding

Both sexes of tsetse fly feed exclusively on the blood of living animals. The insect seldom attacks its host in a single swoop, but usually probes several areas on the animal. Having selected a suitable place, it pierces the skin of the host and, by means of its well-adapted mouthparts, causes small subcutaneous haemorrhages into which it injects an anticoagulant saliva. The insect is then able to suck up the blood from the pool of blood it has made. Once the skin is pierced, the whole feeding process is usually completed in less than a minute.

The fly's weight is often doubled by a complete blood meal. It therefore excretes water and waste products through the anal orifice very soon after feeding, or often during the process. This reduces its weight, and allows it to take off and fly to its resting place. The amount of blood taken depends on several factors. Undisturbed, the fly may take its own weight of blood, or more, most commonly between 15 and 50 mg; but the size of meal depends on climatic conditions, the state of hunger, the species, age and sex of the fly, and, in female flies, the state of pregnancy. These factors may also influence the frequency of feeding, which usually takes place about every 3–4 days. Fertilized females require three or four blood meals for each gestation. Thus, on the basis of an average lifetime of 6 weeks, a female tsetse fly would require about 10–12 feeds, and would take in a total of 150–200 mg of blood.

After feeding, the blood immediately passes to the gut of the fly and causes a visible external distension of the abdomen (figure 1). On the basis of the external appearance and colour of the abdomen, four stages of the digestion, called 'hunger stages', are recognized, from that of the initially engorged fly to that in which no blood remains undigested.

The discovery of, and attraction to, the host

Both sight and smell seem to guide tsetse flies to their hosts, and it is not known which sense plays the greater part. T. A. M. Nash released *Glossina morsitans* into the bush after covering the antennae of the flies with varnish, and he subsequently recovered proportionally as many flies with varnished antennae from the hosts as he did those with unvarnished antennae. This suggested they were able to discover their hosts by sight alone. Experiments with mobile bait-traps indicate that the flies readily perceive moving objects such as motor cars, screens, man, or domestic animals, but are unable to find these objects with untreated antennae and occluded eyes. Usually a fly strikes at the lower parts of the animal host and alights on the underside of the belly, and it mostly favours the legs when attacking man. Hence many traps roughly resemble cattle in shape and size and attract a large number of hungry flies. The situation and the shape of such traps affect the number and the species of flies attracted. Flies are more attracted to traps when animals are in the vicinity or when compounds with particular animal smells are applied to the trap. Thus, W. P. Langridge

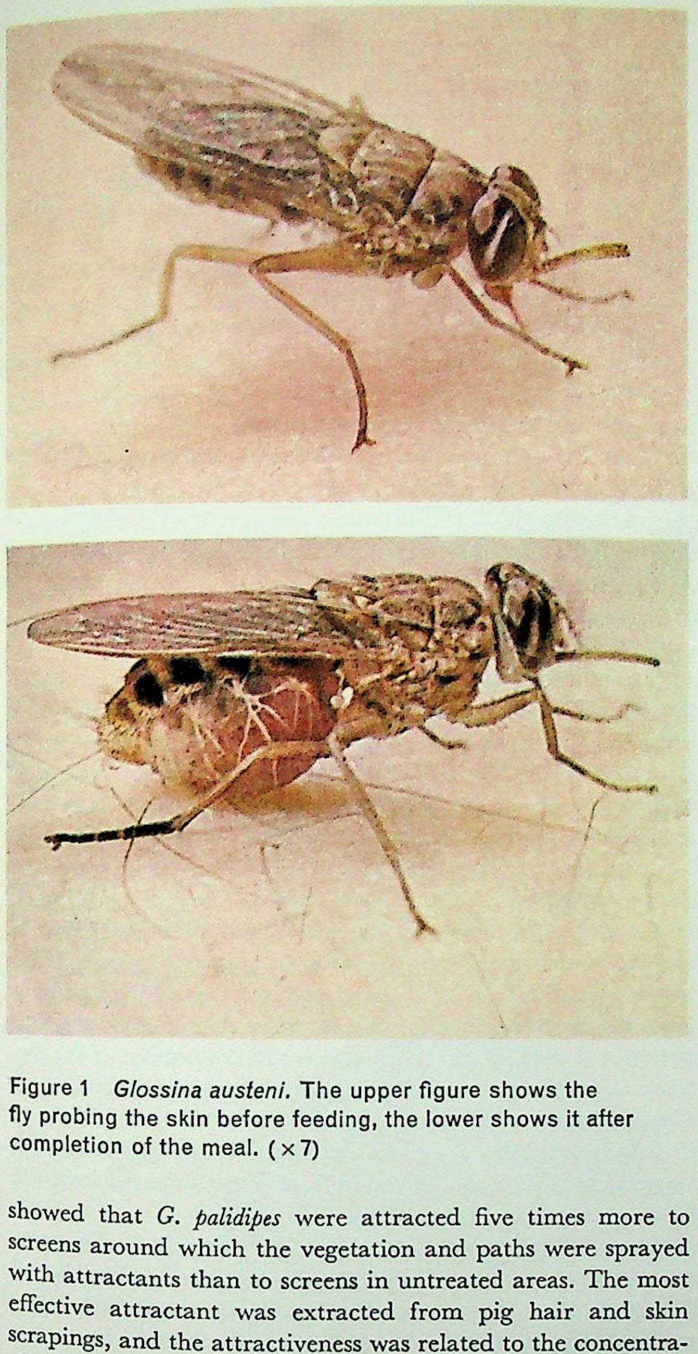


Figure 1 *Glossina austeni.* The upper figure shows the fly probing the skin before feeding, the lower shows it after completion of the meal. (×7)

showed that *G. palidipes* were attracted five times more to screens around which the vegetation and paths were sprayed with attractants than to screens in untreated areas. The most effective attractant was extracted from pig hair and skin scrapings, and the attractiveness was related to the concentration of the extracts. Once the fly has found its host, it explores the skin by smell and probably by touch, more than by sight, before probing and piercing the skin.

C. H. N. Jackson has shown [2] that flies of both sexes visit certain parts of the habitat where they are likely to meet game animals, and that, as hunger increases their activity, they seek the more open spaces. These 'feeding grounds' are open glades, water holes, or areas where the game tracks meet and where the vegetation has been destroyed, and it has been suggested that active flies find these areas when they are hungry as a result of a positive response to light. Thus environmental factors probably influence the choice of hosts, although they do not appear to determine when one species of host is preferred to another.

The natural feeding habits of *Glossina*

In order to determine the natural feeding habits of insects, it is necessary to identify the stomach contents of individual insects from a fully representative and valid sample of the population under study. Unfortunately, the sites where fed insects are captured may influence the pattern of the sample. Such considerations are all-important in estimating the comparative attractiveness to the insect of different hosts. The attraction of flies to a tethered host, such as an ox, indicates the willingness of the insect to feed on the host, but under such conditions the biting incidence does not necessarily represent the natural feeding rate on that host. However, such information is often useful in relation to the transmission of disease by insect vectors. For example, the rate at which human beings in villages are bitten by tsetse flies capable of transmitting human sleeping sickness indicates the extent of potential infection, although the information is valueless in appreciating the natural habits of these flies or in indicating to what extent their usual feeding habits may be altered by the presence of man.

The collection of fully fed flies representing an unbiased sample of the total population demands a knowledge of the resting habits of the insect after feeding. Before these were known, it was possible only to attract flies, irrespective of whether they were hungry or not, to moving screens or other non-living baits, and then select from the large numbers of flies caught those that contained blood-meal residues. Under average conditions, 16 per cent of all the flies of a given population would be expected to contain residual blood in the stomach. In fact, only 1 or 2 per cent of those caught in this way contained sufficient blood for identification, because the flies attracted to screen baits or traps were mostly hungry and hunting for food. Moreover, the samples contained only 1 per cent of female flies, instead of the 50 per cent expected in a normal random population. The natural feeding habits could be determined from such samples only on the assumption that the diet of male and female flies was similar.

After F. Isherwood had described [3] the resting sites of *G. swynnertoni*, it became possible to collect the gorged flies of this species, and later of others, in large numbers within a short time of feeding. These collections contained both sexes in equal numbers and were a more representative sample of the populations under study. Moreover, it became easier to identify the blood, because the contents of the stomach were much fresher than those obtained from hungry flies. The results of identification tests could then be correlated with those from the earlier samples collected by baiting. It was found that, although the earlier identifications of hungry male flies were derived from an admittedly biased sample, the samples corresponded to the feeding habits of engorged male and female flies obtained from resting places.

The methods of identifying blood meals

Any distinctive character of the blood may be used as a means of identification, so that the choice depends on the sensitivity and specificity of the technique. The techniques chosen must be capable of routine use with minimum amounts of material. The diameter of erythrocytes of the blood was used by earlier workers to identify the species of origin, but since the red blood cells of many species are of similar size, identification on this basis is insufficiently specific. Moreover, it is applicable only to very recent blood meals, before the blood cells are altered by digestion. The proteins of the blood plasma, on the other hand, are specific and are more resistant to digestion in the gut of the fly. Their specific nature is determined by their reaction with antibodies.

The precipitin test. In 1910, G. H. F. Nuttall [5] showed that the immunological characters of serum proteins of animals were phylogenetically related. Thus, antibodies prepared in rabbits against the serum protein of an animal react specifically with the proteins of that species and, to a lesser extent, with those from a closely related species of the animal. When the antibody and the corresponding serum protein react *in vitro*, an insoluble compound is formed which can be seen as a whitish precipitate. This 'precipitin reaction' has been used for many years by forensic workers for the identification of human or animal blood stains. Since precipitates occur even with very small concentrations of the serum proteins, the method is ideal for the identification of blood meals, which as a result of digestion often contain only a small residue of the proteins originally ingested by the insect.

The reaction is allowed to take place in a narrow tube by superimposing the blood-meal extract on a layer of antiserum. Because of the higher specific gravity of the antiserum, mixing of the reagents is limited, and the precipitins occur at the junction of the two reagents as a whitish ring. The density of the precipitin 'ring' at the interface indicates the concentration of the serum proteins in the blood meal. The technique can be readily automated, and a specially designed apparatus enables a single operator to perform a very large number of tests in a short time. However, even with carefully prepared antisera, the value of the method is limited, because it is insufficiently specific in certain instances [6, 7]. For example, the precipitin reaction can easily distinguish goat from ox, but cannot distinguish goat from sheep.

The inhibition test. A more discriminating test was based on Boyden's serological technique for qualitative determination of tuberculins. In this 'inhibition test', the erythrocytes of normal rabbits are coated with tannic acid, and treatment with the serum of an animal coats them with the serum proteins. In this state they will be agglutinated by an antiserum prepared against the serum proteins of that species of animal. If the antiserum, at a suitable concentration, has previously been mixed with the blood-meal extract originating from that animal, the specific antibodies will have been neutralized and the coated cells no longer agglutinate. A blood meal is identified by its ability specifically to prevent agglutination of red blood cells sensitized with the serum of a given species of animal. This inhibition test, although very specific and extremely sensitive, is much more complicated and laborious than the precipitin test.

The blood meals are first screened by the precipitin test to find out which group of animals—primates, suids, bovids, carnivores, for example—they come from. The individual species is then identified by the inhibition test. Thus the primate feeds are tested for the presence of human, baboon, and monkey blood, and suid feeds for warthog, bushpig, and giant forest hog. Bovid feeds may have to be tested for ten or more different species of ruminants before they can be identified. Animals such as elephant, hippopotamus, rhinoceros, and porcupine can be identified directly by precipitin tests.

Validity of methods

The results of the identification tests were checked by tests with known blood meals. In view of the differences in the rate of metabolism of tsetse flies in the wild state and those kept under captivity, the checks were made under conditions resembling those in the field as closely as possible. A number of wild hungry *G. morsitans* were fed artificially on man, donkey, domestic pig, and captive gazelles on different days, and marked with paint on the back of the thorax. The mark indicated the date of marking and the identity of the individual fly. The marked flies were released in the bush and then recaptured, together with a number of unmarked flies, on subsequent days. All blood meals were then identified without any knowledge of the identity of the flies. Of 500 flies fed and marked in this way, 28 were recovered. All the meals were identified correctly, which showed that the technique was valid even after digestion had proceeded in the wild state up to a day or more [4]. Numerous other control tests with captured flies kept in captivity after artificial feeding confirmed the validity of the identification techniques.

Wild hosts of tsetse flies

A study of the hosts of fifteen species of tsetse flies (table 1) from many different areas in East, West, and Central Africa was based on 22 000 blood-meal identifications [8]. The main hosts were present in most areas, although the density of various species of game varied considerably. In most places it was known only that particular animal species were present or absent, but more exact proportions of different host species were assessed in selected areas, so that the availability of hosts to tsetse could be carefully studied. The relative density of different game species could, even so, be estimated only approximately; it was difficult to relate results of identification tests to the numbers of hosts available. However, the feeding habits of species of tsetse were generally consistent in most collecting areas, and the patterns for different species of *Glossina* could be averaged for all areas.

The pattern of feeding was characteristic for each species of tsetse and was broadly divided into five main categories according to the main hosts mostly used. In the first group (figure 2), tsetse flies fed mostly on suids. Different species of suids were present in different areas, and the local tsetse flies fed indiscriminately on whatever species was available. Thus, in Tanganyika, *G. swynnertoni* mostly fed on warthog, while in Zanzibar, *G. austeni* fed on the bushpig, the only pig in the area. In the West African forests 70 per cent of *G. tabaniformis* fed on the red river hog, and in Uganda *G. fuscipleuris* derived the larger part of its food from the giant forest hog. The remaining hosts of the tsetse flies in this category largely reflected the presence of the animals in the area or in the habitat of the fly. Thus *G. fuscipleuris* derived nearly 15 per cent of its feeds from hippopotamus, which corresponds to the extent to which both fly and host inhabit the riverside. Similarly, in Nigeria, about 20 per cent of all feeds of *G. tabaniformis* were obtained from the forest porcupine.

A second group of flies (figure 3) comprised the three savannah subspecies of *G. morsitans*. All fed mainly on suids and bovids in about equal proportions. The bovid feeds were mainly from buffalo, kudu, eland, and bushbuck, but other bovid hosts, such as giraffe and roan antelope, were also used.

Table I
Sources of blood meals of *Glossina.*

G. palpalis palpalis	Nigeria, Liberia
G. palpalis fuscipes	Tanganyika, Kenya, Uganda, Nigeria
G. tachinoides	Nigeria
G. morsitans morsitans	Tanganyika
G. morsitans orientalis	Tanganyika, Kenya, Northern Rhodesia, Southern Rhodesia
G. morsitans submorsitans	Nigeria, Uganda, Sudan
G. swynnertoni	Tanganyika, Kenya
G. pallidipes	Tanganyika, Uganda, Kenya
G. longipalpis	Nigeria
G. brevipalpis	Uganda, Kenya
G. longipennis	Kenya
G. austeni	Kenya, Zanzibar
G. tabaniformis	Nigeria
G. fuscipleuris	Nigeria, Uganda
G. fusca	Nigeria

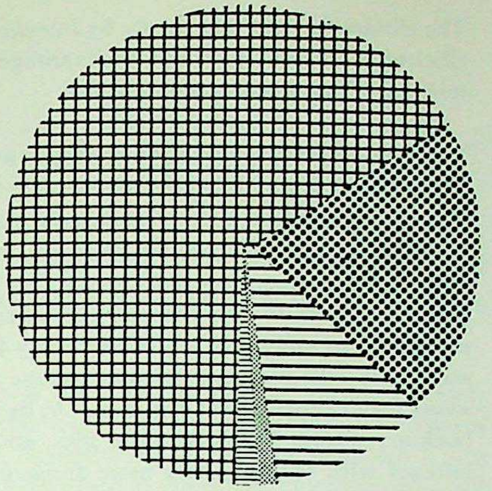


Figure 2

G. swynnertoni

G. austeni

G. tabaniformis

G. fuscipleuris

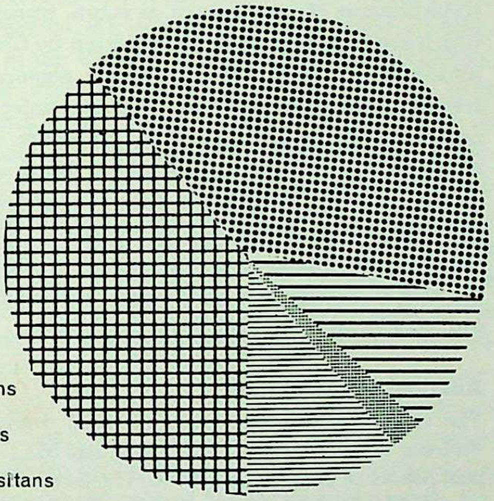


Figure 3

G. morsitans morsitans

G. morsitans orientalis

G. morsitans submorsitans

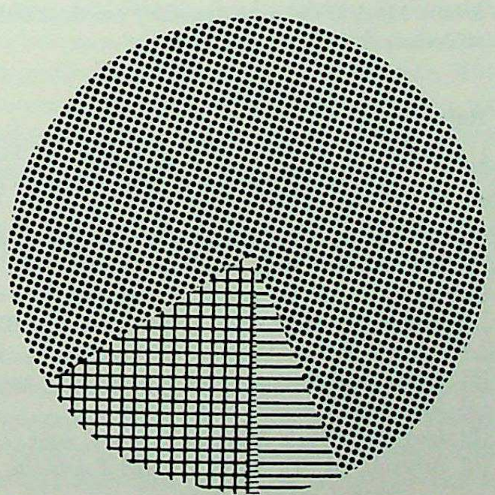


Figure 4

G. pallidipes

G. longipalpis

G. fusca

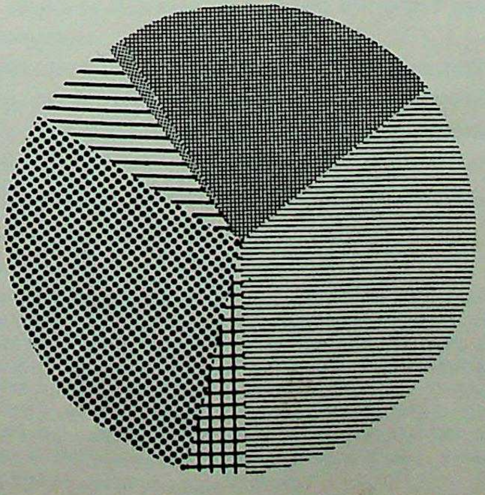


Figure 5

G. palpalis palpalis

G. palpalis fuscipes

G. tachinoides

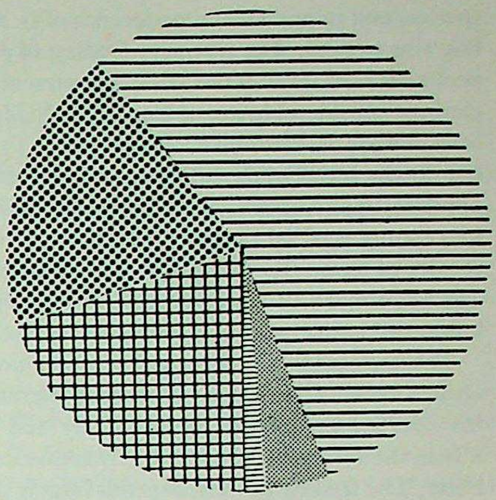


Figure 6

G. longipennis

G. brevipalpis

Primates	▤	Suids	▦
Bovids	▩	Mammals	☰
Birds	░	Reptiles	▓

Figures 2–6 Pattern of feeding of various tsetse-fly species.

The third feeding pattern (figure 4) was represented by *G. pallidipes*, *G. longipalpis*, and *G. fusca*, all feeding mainly on bushbuck. *G. pallidipes*, which is limited to East Africa, had the same feeding habits as the West African species *G. longipalpis*, to which it corresponds bionomically and ecologically. The bushbuck is the only wild host from which the human sleeping sickness organism (*Trypanosoma rhodesiense*) has been recovered, which justifies the inclusion of *G. pallidipes* among the vectors of Rhodesian sleeping sickness.

The fourth category (figure 5) of feeding patterns contains the three species of tsetse fly—*G. palpalis fuscipes*, *G. p. palpalis*, and *G. tachinoides*—that are mainly concerned with the transmission of Gambian sleeping sickness in man. These are very similar flies, and their habitat is restricted to the edge of watercourses and lakes. The feeding pattern of these species was extremely variable and depended largely on the availability of hosts. Unlike other species of *Glossina*, these flies are consistently attracted to man, particularly in villages situated along rivers; fishermen are also constantly attacked by the *palpalis* group of tsetse fly. In undisturbed areas, aquatic reptiles such as the crocodile and monitor lizards were the main hosts, and *G. palpalis* maintained itself entirely on these reptiles in some places. Bushbuck, which come to water in the habitat of these flies, are particularly attacked, but it seems that any mammalian host may be bitten. Local differences in feeding patterns therefore occur.

Finally (figure 6), the two largest tsetse flies, *G. longipennis* and *G. brevipalpis*, have very characteristic feeding habits. Most of the time, *G. longipennis* feeds on rhinoceros, but the migration of elephants and buffalo into the area diverted the fly to these large hosts. *G. brevipalis* attacked hippopotamus in most areas, but an equal proportion of its food in some areas was obtained from bushpig.

Feeding habits and the ecology of tsetse flies

Clearly, some animals are used more frequently than others by most tsetse flies. The suids appear to be a favourite host of most species, and in one area, a preference of *G. swynnertoni* for warthog was proved. The intensive feeding of some species on the preferred host, even when these are present in small numbers, causes a considerable loss of blood in individual animals. In one area, individual bushbuck lost as much as 17 grams of blood per day by the biting of *G. pallidipes*, and in another, a warthog lost between 13 and 27 grams per day to *G. swynnertoni*. In contrast, vast numbers of zebra and wildebeest are never bitten by any species of tsetse fly, and hosts such as waterbuck, hartebeest, and impala, all of which occur in very large numbers in the habitat of many species of tsetse flies, are seldom bitten.

The reasons for this very marked discrimination are not at all clear. Certainly, in many areas the preferences are not due to the overlapping of the habitat of fly and host, although in others the choice of hosts is clearly influenced by their availability. The feeding pattern depends largely on the adaptability of the fly to its environment, and some species of *Glossina* may more readily change their feeding habits than others. For example, *G. morsitans*, which normally prefers suids and bovids, may readily change to other hosts in the absence of its favourite diet. Gorged *G. morsitans* collected in an area in Ankole, Uganda, showed the usual feeding pattern. 67 per cent had fed on warthog, 7 per cent on bovids. In a similar area only 20 miles away, where hippopotamus was present, 20 per cent had fed on the warthog and 25 per cent on hippopotamus. Kudu replaced warthog as the source of food in a place in Tanganyika where there were a few suids only. The opportunist feeding pattern of the *palpalis* group of flies also indicates a certain adaptability to the availability of hosts.

Control of tsetse by starvation

The idea of controlling or eliminating tsetse flies in selected areas by the elimination of their favourite food is attractive. The success of such methods would, of course, depend on the degree and rate of adaptation of the fly to a new diet. Flies such as those of the *palpalis* and *morsitans* group probably have too wide a range of possible hosts to suffer greatly by the selective elimination of their preferred hosts. However, if elimination of fly by game control is used, the species of animals that are never attacked by tsetse fly could reasonably be left without much danger of providing new sources of food. Thus if all game were destroyed except zebra, waterbuck, wildebeest, impala, hartebeest, and most small mammals, tsetse fly would probably starve to death. Other tsetse flies such as *G. pallidipes*, *G. longipalpis*, *G. longipennis*, and *G. fusca* might suffer more severely by the removal of the few hosts they normally preferentially feed on. However, until suitable trials can be made and economical methods of destroying particular hosts can be devised, the success of such procedures must remain conjectural.

The elimination of the tsetse fly by insecticide spraying is very efficient, but it has the serious disadvantage of leaving the main reservoir of trypanosomes untouched.

The food of tsetse in relation to trypanosomal infections

The absence of any obvious correlation between the incidence of trypanosomal infections in a species of animal and the degree to which it is used by tsetse fly raises important biological issues [9]. The frequency of biting by flies and the consequent heavy exposure to trypanosomal infection does not appear necessarily to initiate an active immunity against infections in such hosts—the resistance of many game species to trypanosomes may be innate, not acquired. On the other hand, waterbuck are seldom bitten by tsetse flies, and these animals are infected with trypanosomes more frequently than any other game species. The comparatively high rate of infection with trypanosomes of species such as zebra, impala, and hartebeest, which are infrequently or never bitten by *Glossina*, suggests that infections are transmitted by other vectors. These might be tabanid flies or *haematopota* carrying infected blood on the proboscis. This mechanical transmission may be the more common one in infections within a herd, whereas the widespread transmission between herds or between different species of animals is more likely to be mediated by *Glossina*. The presence in a given area of several species of *Glossina* with different yet partly overlapping host preferences will also promote the interspecies spread of infection in animals.

Acknowledgments

The colour photographs were obtained through the courtesy of Professor W. Kershaw, Dr W. Crewe, and Mr J. Brady of the Liverpool School of Tropical Medicine. The flies in the photographs were provided by Dr T. A. M. Nash, Tsetse Fly Laboratory, University of Bristol. Much of the work described was financed by the Department of Technical Co-operation.

References

[1] Buxton, P. A. 'The Natural History of Tsetse Flies'. London School of Hygiene and Tropical Medicine, Memoir No. 10, 816 pp. 1955.

[2] Jackson, C. H. N. *Bull. ent. Res.*, **24**, 443, 1933.

[3] Isherwood, F. *Ibid.*, **48**, 601, 1957.

[4] Weitz, B. and Buxton, P. A. *Ibid.*, **46**, 531, 1955.

[5] Nuttall, G. H. F. 'Blood Immunity and Blood Relationships'. Cambridge University Press, London. 1904.

[6] Weitz, B. *Bull. World Hlth Org.*, **15**, 473, 1956.

[7] *Idem. Exp. Parasitol.*, **9**, 63, 1960.

[8] *Idem. Bull. World Hlth Org.*, **28**, 711, 1963.

[9] Ashcroft, M. T., Burtt, E., and Fairbairn, H. *Ann. trop. Med. Parasit.*, **53**, 147, 1959.

Geophysical studies in the Indian Ocean

B. D. Loncarevic

The Indian Ocean is important to the densely populated countries that surround it, but it has been very little investigated. An International Indian Ocean Expedition is now in progress, and this article is intended to provide a background to the accounts of that expedition. It describes the results of the expedition of H.M.S. *Owen* to the Indian Ocean in 1961, a preliminary to the main British expedition, and discusses them, particularly in the light of the drifting-continent theory.

The Indian Ocean is bordered by some of the most densely populated areas in the world and crossed by the oldest routes of oceanic commerce. Nevertheless, up to a few years ago it was the least known of the oceans of the world. Oceanography had bypassed it, and the question of the origin, characteristics, and effects of the monsoons, for example, were as much a puzzle as they had been to the ancient Arab mariners in their dhows. To solve some of these problems, an expedition has been organized, involving some 30–40 ships from ten nations, as an international contribution to the development of the countries bordering the Indian Ocean. This will study the large-scale interactions between wind and sea, the equatorial currents, the productivity of the ocean, and the structure and composition of the ocean bottom, and will provide data for many years of research and interpretation. This article describes recent work in the Indian Ocean, and is intended to give a background for the forthcoming reports of the International Indian Ocean Expedition.

In 1961, H.M.S. *Owen*, one of Her Majesty's surveying ships, was given the task of carrying out a reconnaissance survey of the area bounded by Africa on the west, Pakistan and India on the north and east, and the latitude of Mauritius on the south. The Department of Geodesy and Geophysics, Cambridge University, supplied scientific instruments and personnel and undertook the reduction and preliminary interpretation of the data. The purpose of the reconnaissance expedition was to outline the major features, and to delineate areas for future detailed studies by the British ships of the International Indian Ocean Expedition. The results of the survey have been published by the Hydrographic Department of the Admiralty [1] and form the basis of the following discussion. Figure 1 shows the ship's track.

Instrumentation

The most important instrument used was a precision echo-sounder, designed and built by A. S. Laughton at the National Institute of Oceanography. This instrument used an outboard transducer, and the time required for the echo to return from the ocean bottom was measured accurately with reference to the constant frequency of a tuning fork. Changes in depth of one to two metres could be detected even when the depth was more than five kilometres.

The second instrument in continuous use was a magnetometer developed by M. N. Hill at Cambridge University and built commercially in Edinburgh. This instrument measures the proton-precession frequency, which is a function of the local magnetic field, with an accuracy of one part in 100 000. The sensor, a bottle of water with a coil wound round it, was towed 130 metres behind the ship to minimize the distortion of the Earth's field by the presence of the ship. The associated electronic apparatus, chart recorder, and punched-tape output device were installed in the chart room.

The third instrument used routinely throughout the survey was an Askania-Graf sea gravimeter. This instrument is a very sensitive spring balance, highly damped in order to eliminate the effect of heaving of the ship, and mounted on a gyro-stabilized platform to compensate for roll and pitch of the ship. A gravimeter can measure with great precision the variations, due to the variations in the shape and structure of the ocean bottom, of the gravitational attraction. However, because of the residual effect of ship motions and inherent errors in navigation by star fixes alone, the overall accuracy of the gravity measurements was only of the order of one part in 100 000. Instruments for precise measurement of depth, of magnetic and gravitational fields, and of the sediment thickness

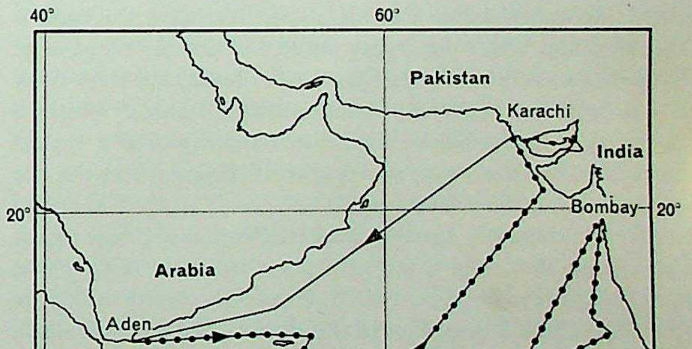


Figure 1 Tracks of H.M.S. *Owen* in the Indian Ocean during the 1961–62 surveying season. Dots show the places where observations were made.

B. D. Loncarevic, *M.A., Ph.D.*
Was born in 1930 in Belgrade, Yugoslavia. After graduation from the University of Toronto, he worked with an airborne geophysical unit on the development of an electromagnetic prospecting instrument. In 1958, he went to Cambridge on an Athlone Fellowship to work at the Department of Geodesy and Geophysics on the problem of gravity measurement at sea. During 1961–62 he was a member of the U.K. scientific team taking part in the International Indian Ocean Expedition. He is now a Senior Scientific Officer at the Bedford Institute of Oceanography, Dartmouth, Nova Scotia, where he is in charge of the marine geophysical programme.

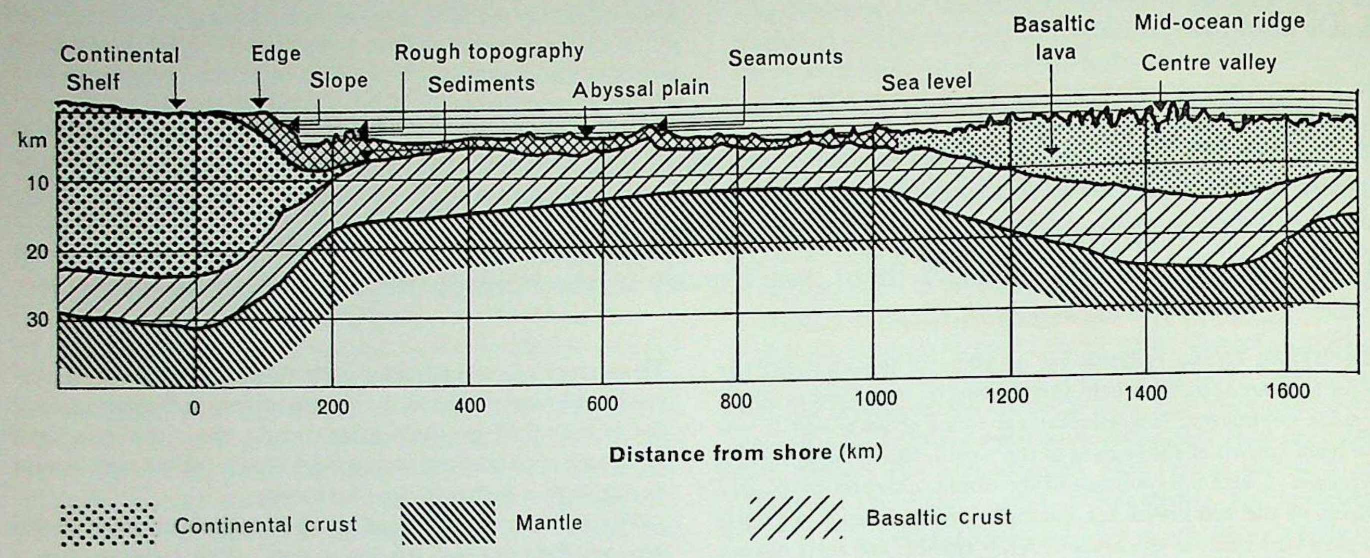


Figure 2 A diagrammatic cross-section through the continental and oceanic crust.

are indispensable tools in the study of the characteristic features of the ocean bottom.

It is only in the last ten or twenty years that the fascinating variety of the ocean-bottom features has been recognized. The crust, which is the outer layer of the solid Earth overlying the more dense and more uniform layer called the mantle, has a different structure and composition under the oceans from those under the continents. The continents are composed of granitic rocks of about 35 km thickness, though this increases to 60 km under some mountain ranges. The oceanic crust is made up of sediments underlain by basaltic rocks. The total thickness is about 10 km, with an increase under the oceanic mountains. Granitic rocks are normally completely absent, although there is one notable exception, the Seychelles Islands, which is discussed at length below. Figure 2 is a diagram of a typical cross-section of the crust, showing the difference between the continents and the oceans.

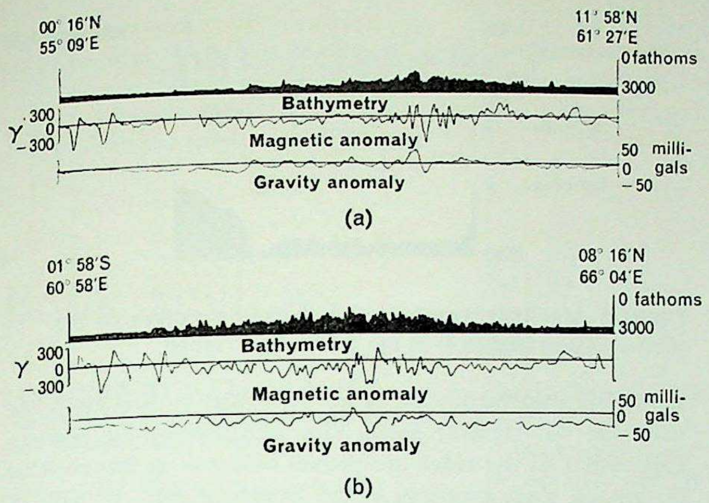
All the oceans are broadly similar. Beyond the edge of the continental shelf, the depth of the sea increases rapidly and some rough topography at the base of the continental slope leads to a plain. The surface of this abyssal plain, which slopes away from the continents with a gradient of less than one part in 1000, consists of slowly deposited sediments; it is not perfectly smooth, because of burrowing animals, ripple marks, and an occasional river channel or mid-ocean canyon. Abyssal hills, rising a couple of hundred metres, appear scattered on the plain; if they reach a height of over 1000 m they are called seamounts. As one moves away from the continents towards the middle of the ocean, the topography becomes rougher, until a mountain ridge is encountered with rugged peaks and deep valleys. This ridge follows the median line of the oceans round the world and represents the longest mountain chain on the Earth. Occasionally, the mountain peaks reach above the sea surface, giving rise to islands such as the Azores. Islands also occur away from the ridge, either isolated, as, for example, Madeira, or in chains. Of special interest are island arcs, which are curved chains of islands with a deep trench and a large negative gravity anomaly on the convex side of the arc.

Most of the major ocean-bottom physiographical features are present in the north-western Indian Ocean. There are two fully developed abyssal plains, and abyssal hills that merge with the mid-ocean ridge. The ridge is a branch of the world-wide system of mid-ocean ridges and may have a continental extension at its northern end. The area is studded with islands and coral atolls, and there are seamounts and submarine banks and even a deep-sea trench. The area is an attractive one for geophysical studies, and it is hoped that some major advances of knowledge will come from further investigations.

The Carlsberg Ridge

The major physiographic feature in the area is the Carlsberg Ridge, an impressive chain of mountains over 4000 km long. It emerges from a complicated and little known area of rough topography somewhere south of Mauritius and stretches to the entrance of the Gulf of Aden. It was first recognized on a crossing by the 'Dana' expedition in 1930 between latitudes 1° and 2°N, and on the basis of this single crossing Johannes Schmidt postulated the existence of a ridge and gave it the name 'Carlsberg' after the sponsor of the expedition. It was further investigated by the 'John Murray' expedition in 1934 [2] and has been crossed by many expeditions since then. These crossings have established the continuity of the ridge and some of its major characteristics. The abyssal plains in the north are well developed right up to the flanks of the ridge. Its width on the two crossings in the northern section is 750 km and 950 km and the peak is at 1850 m, thus rising 2350 m from the level of the abyssal plain. The sections of the ridge at A and B (figure 1) are illustrated with the profiles of bathymetric, magnetic, and gravitational anomalies in figure 3. The most striking feature of these profiles is the appearance of a deep valley near the crest of the ridge. At A the north wall is over 2200 m high, the south wall is about 1500 m high, and the valley is 20 km wide. At B, both sides of the valley are about 1500 m high and at 500 m elevation the valley is 12 km wide. On both crossings the depth to the floor of the valley is 4200 m and represents the greatest depth within the central part of the ridge. Although the valley seems to be a characteristic of the ridge, its continuity throughout the length of the ridge is questionable. Detailed investigations around 5°N and 62°E have shown that the valley is disrupted there by fractures and faults. It seems likely that the valley has been disrupted, blocked, and twisted in various places, but at the same time it remains a basic characteristic of the ridge system and deserves more detailed study.

The magnetic anomaly (figure 3) across the ridge shows a disturbed field with generally poor correlation between the magnetic and bathymetric profiles. The only place where there is a clear correlation between the two profiles is over the central valley itself. A distinct negative anomaly equivalent to



Figures 3a and 3b Profiles illustrating the topography of the Carlsberg Ridge and the associated magnetic and gravity fields. (1 gal=1 cm/sec^2; 1 $\gamma=10^{-5}$ oersted)

more than 1 per cent of the local magnetic field was measured when crossing the valley.

It is impossible to say from the magnetic profile what causes the anomaly. However, it is possible to calculate magnetic anomalies over assumed bodies of varying shape, size, and magnetic contrast, and to compare these calculated values with the observed one, thus obtaining some general idea about the possible source of the observed anomaly. These calculations have been made, taking into account the strength and inclination of the local magnetic field and the general direction of orientation of the ridge. It was found that a rectangular prismatic body of some 5–10 km cross-section, very long compared with its cross-section and only a few kilometres under the valley, would give an anomaly corresponding fairly closely in shape and magnitude with the observed one. Similar results have been obtained in the Atlantic ocean, around 45°N, and both there and in the Indian Ocean, they can be explained by assuming that there is a dyke intrusion underlying the valley floor. This would also conveniently explain the high heat-flows observed in the centre of the ridge. This provides further evidence that the chains of mountains on the ocean floor are a part of a world-wide system, and that the same mechanism was responsible for all of them.

There are no other anomalies on the ridge that compare with the anomalies over the valley. However, on the smooth abyssal plains bordering the ridge similar large magnetic anomalies were encountered, and these were completely unexpected. The type and shape of the magnetic anomalies under the abyssal plains on either side of the Carlsberg Ridge can be seen on the south (left) end of the profiles in figure 3; they were even more marked and persistent north of the ridge. There were no visible seamounts associated with magnetic anomalies on the flanks of the Carlsberg Ridge, and it is extremely unlikely that there were seamounts hidden under the sediments, as they would then have been detected on the gravimeter trace. The gravimeter trace on the abyssal plain remained very steady and indicated that under these plains were uniform layers. The origin of these magnetic anomalies is still a mystery, but it is possible to guess the depth of a body that could cause these anomalies by assuming that the depth is roughly equal to the 'half-width' of the anomaly. Using this rule, the approximate depth of the body causing the anomaly is 25 km. Thus, it seems that the cause of these anomalies is in the upper mantle, below the base of the crust.

Two hypotheses can be put forward to explain the abyssal plain anomalies, but it will be some time before they can be evaluated. The first one derived from the obvious similarity in shape and magnitude between the abyssal plain anomalies and those under the median valley. If convective motion is assumed in the hot, relatively plastic mantle, then a volcanic intrusion and its associated magnetic anomaly might be due to the material brought up by a convection current. This convection current would rise under the median valley and cause the magnetic anomaly over it. While coming up, this material would be hot and non-magnetic, but after it cooled down below its Curie point it would become magnetized and the magnetization would become 'frozen' into the intrusion. The nature and mechanism of the convection current motion is not well understood, but, according to one hypothesis, there is a period of rest after the current completes one cycle. During this period of quiescence, the intrusion might be cooling down and acquiring the frozen-in magnetic anomaly. During the next cycle of motion of the convection current, new material is brought up from the depths and material brought up in the previous cycle is pushed along the base of the crust. It eventually disappears under the continents and sinks to the depths, thus replacing the material rising under the ridge. The movement under the crust carries along the frozen-in anomalies. They appear as distinct anomalies because of the postulated pulsating nature of the convection currents. The movement envisaged would be intermittent and the distance travelled during the mobile phase would be represented by the distance between the anomalies.

The hypothesis in the above form has many points that could be criticized, but it does fit in with recent ideas of H. Hess and R. Dietz [3] about the spreading of the ocean floor. It also fits in with some ideas of J. Tuzo Wilson [4], who collected all the known age determinations from the oceanic islands and showed that the ages increase with distance from a ridge. This is presumably because the older islands have travelled further. Thus the postulated motion of the upper mantle could be the driving force for the crustal movements.

The second hypothesis explaining the magnetic anomalies under the abyssal plains is also based on the concept of convection currents, but does not involve the median valley anomaly. It is again assumed that the convection current is rising under the ridge, flowing under the crust away from the ridge, and sinking under the continents. The upper boundary of this flow might not be smooth, indeed it is more likely to be undulating, so that the depth of the isothermals is variable. This would imply a variable depth of materials, at temperatures below the Curie point, which could acquire a magnetization that would manifest itself as the observed magnetic anomaly. The density differences between material above and below Curie point would be too small to be detected with the present equipment for measuring gravity at sea. On the other hand, the undulations of the upper boundary of the convection current could account for some of the observed variations of the heat flow. Indeed, a dense profile of heat-flow stations might be a way of testing this hypothesis.

Seychelles Bank and Amirantes Trench

To the south and west of the Carlsberg Ridge lies the Seychelles Bank with about 100 islands, 24 of which are inhabited. While most of the outlying islands are low coral islands, the central part of the Mahé-Praslin group is granitic in structure and over 600 million years old (Precambrian) [5]. It is the only known mid-ocean island group composed of granite, and a special explanation is required for its origin and present location, as granitic rocks have always been considered as

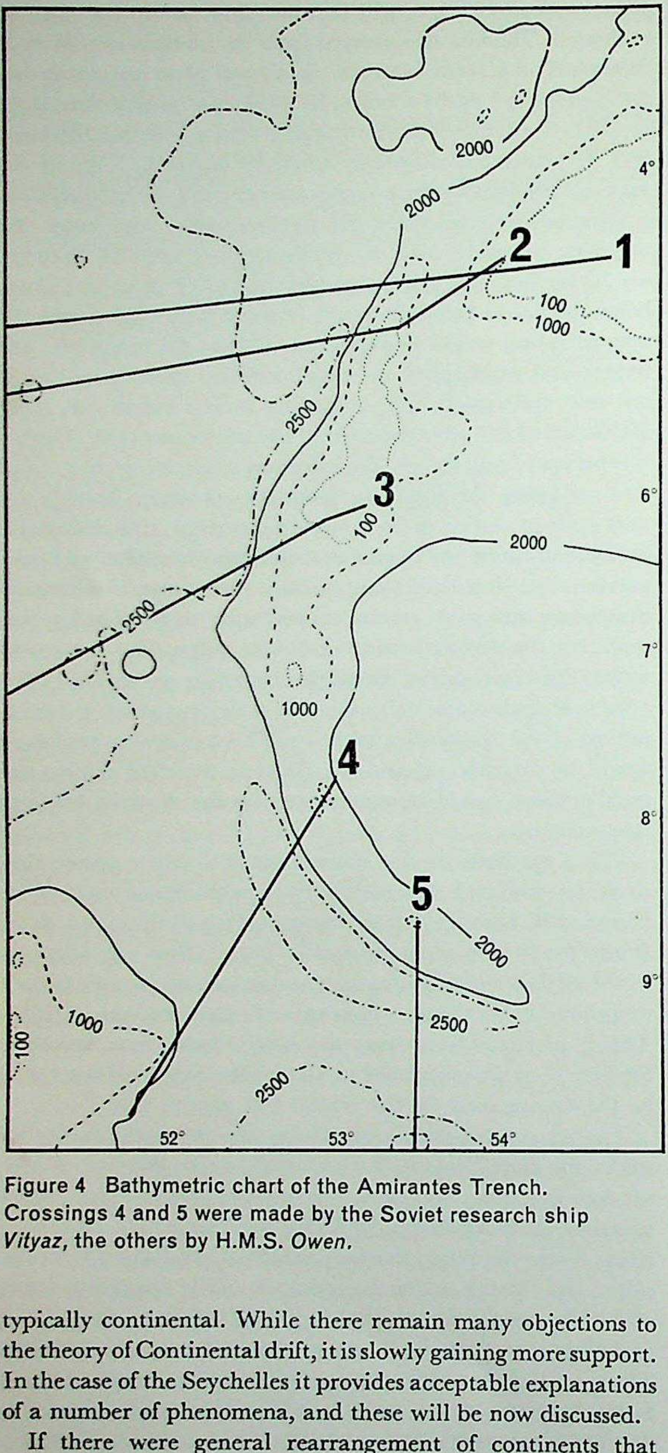


Figure 4 Bathymetric chart of the Amirantes Trench. Crossings 4 and 5 were made by the Soviet research ship *Vityaz*, the others by H.M.S. *Owen*.

typically continental. While there remain many objections to the theory of Continental drift, it is slowly gaining more support. In the case of the Seychelles it provides acceptable explanations of a number of phenomena, and these will be now discussed.

If there were general rearrangement of continents that opened up the Indian Ocean as we know it today, it is natural to suggest that the continental part of the Seychelles has dropped off and was left behind, but where did it belong originally? If the Carlsberg Ridge divides the Indian Ocean, the Seychelles could reasonably be classified as part of the original African continent. The other possible connection is via Madagascar, and even a cursory glance at the map of the Indian Ocean will show the remarkable straightness of its east coast, which points towards the Seychelles. A large linear fault has been mapped on this coast, and it was only natural that a northward drift along this fault has been postulated for the Seychelles, despite the paucity of evidence. The evidence is still not conclusive. But new facts have emerged that support the idea. Seychelles Bank is bounded on the west side by a low-lying ridge of coral islands called the Amirantes. Repeated crossings with a shipborne magnetometer have indicated that the ridge is underlain by highly magnetic rocks rising close to the surface underneath the coral. This ridge is directly in line with the Madagascar fault. Five crossings of the western approaches to the ridge are shown in figure 4. The striking feature on these crossings is the trench recently noticed by A. S. Laughton and called by him the Amirantes Trench. The gravitational anomaly profile reproduced in figure 5 indicates a uniformly increasing gravitational anomaly on top of which is superimposed a local negative anomaly of $-0{\cdot}12$ cm/sec^2. The magnitude of this anomaly is comparable to that observed over other oceanic trenches, such as the Aleutian Islands Trench.

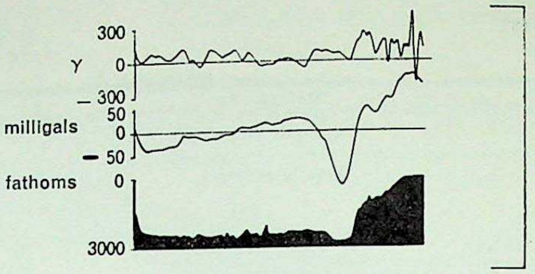


Figure 5 Magnetic, gravity, and bathymetric profiles on the third crossing (figure 4) of the Amirantes Trench.

The deep-ocean trenches have been investigated in other areas, notably the Pacific, and a general conclusion is that they are of tectonic origin. Together with the east coast of the Madagascar Fault, the Amirantes Trench is the most important new evidence for tectonic activity in this area. If the strike-slip origin for this trench could be proved, the riddle of the Seychelles might be solved and their ancient connection with Africa reconstructed by moving them south to the east coast of Madagascar.

If the above were true, it would be natural to expect further evidence of cracking of the Indian Ocean floor to the north of Seychelles. This area has not been systematically investigated because, until now, nobody expected to find anything interesting or unusual there. Further north, where the Carlsberg Ridge approaches the Horn of Africa (Cape Guardafui), D. H. Matthews [6] and his colleagues have recently shown, by skilful synthesis of bathymetric and seismic evidence, that the ridge has been fractured and the western section moved perhaps as much as 300 km to the north. Between Matthews' fault and the Amirantes Trench, the evidence might be buried or confusing. Nevertheless, these two features indicate that there might be an enormous fault on the bottom of the ocean stretching from the southern tip of Madagascar up to the northern shore of the Gulf of Aden and beyond. If this crack could be confirmed, its existence would profoundly affect our thinking about the structure of ocean floors.

Future work

In order to investigate further many fascinating puzzles on the bottom of the Indian Ocean, the new tools and techniques of marine geophysics will have to be applied on a large scale. In particular, seismic surveys of the thickness of various crustal layers will have to be extended. Both American and Russian scientists have conducted some measurements, but there is certainly room for an extension of their work. Many more heat-flow measurements will be required, and they will have to be densely spaced along selected profiles. Palaeomagnetic measurements should be made on rock specimens from Seychelles and Madagascar. But, above all, systematic, detailed surveys and investigations of selected portions of the ocean floor must be carried out in spite of the difficulties, cost, and tedious nature of the work. The progress from wild idea to hypothesis and thence to theory can be based only on factual survey information.

References

[1] Loncarevic, B. D. and Matthews, D. H. 'Bathymetric, magnetic and gravity investigations, H.M.S. *Owen* 1961–1962'. Admiralty Marine Science Publications No. 4, Pts. 1 and 2. London. 1963.

[2] Farquharson, W. I. 'Topography'. The John Murray Expedition, Scientific Reports, Vol. 1, No. 2. London. 1936.

[3] Dietz, R. S. *Nature, Lond.*, **190**, 854, 1961.

[4] Wilson, J. *Sci. Amer.* **208**, 86, 1963.

[5] Baker, B. H. and Miller, J. A. *Nature, Lond.*, **199**, 346, 1963.

[6] Matthews, D. H. *Ibid.*, **198**, 950, 1963.

Book reviews

TEACHING OF MATHEMATICS

The U.S.S.R. Olympiad Problem Book, *by D. O. Shklarsky, N. N. Chentzov, and I. M. Yaglom (revised and edited by I. Sussman, translated by J. Maykovich). Pp. xvi+452. W. H. Freeman & Co., San Francisco and London. 1962. 63s. net.*

This very interesting book covers 320 unusual problems, and their solutions, in algebra, arithmetic, elementary number theory, and trigonometry. The problems are at high-school level—some are advanced—and are taken mostly from mathematical competitions held in Russian schools. The solutions are given in considerable detail and occupy the major part of the book. Teachers of mathematics will find the problems of great interest, especially in view of the world-wide recent modernization of the school mathematics syllabus, and the cultivation, in society generally, of a greater interest in mathematics.

During the last decade or so, there has been a strong trend, particularly in America, to push back into the schools some of the abstract notions of mathematics, together with the appropriate terminology. Schoolchildren will have heard, on many occasions, of sets, operations, mappings, rings, fields, vector spaces, Boolean algebra, and so on. The fashion spreads and carries with it a social obligation to fall into line and follow the crowd. It is a dangerous experiment, and may lead to a generation mathematically illiterate. The Olympiad Problem Book would suggest that the Russians are having none of this. The problems cover the most classical parts of mathematics, and seem to indicate that the best training for the young mind is still to come to grips with problems—some of them rather slippery—whose solution depends on semi-original applications of standard results in elementary mathematics. All teachers concerned with the 'new' thinking in mathematics should study the book carefully.

L. S. Goddard

PROCESS CONTROL

The Mathematical Theory of Optimal Processes, *by L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. Authorized translation from the Russian by K. N. Trirogoff. Editor: L. W. Neustadt. Pp. viii+360. Interscience Publishers, a Division of John Wiley & Sons, New York and London. 1963. 90s. net.*

This book, written by the recipients of the 1962 Lenin Prize for Science and Technology, makes available the powerful results in optimal control theory obtained by the group of mathematicians led by Academician Pontryagin at the Steklov Mathematical Institute in Moscow.

In general terms, the object of this book is to describe the mathematical theory of finding the best method of carrying out the controllable physical processes that take place in modern technology. The concept of the optimal control of the process does, of course, vary from one situation to another. It may refer to rapidity of action or accuracy of achievement, but whatever type of optimum is envisaged, the problem reduces essentially to one in the calculus of variations. The interesting point is that these variational problems almost without exception lie outside the classical calculus of variations. They do, however, receive considerable unification from one general mathematical method called the maximum principle, due essentially to Pontryagin.

After an exposition of the maximum principle and the formal proof, the authors return to applications to linear time-optimal processes, to optimal processes with restricted phase coordinates, and to applications to statistical problems. In addition to this, they give an interesting account of the relation of the maximum principle to the classical calculus of variations and they survey a number of miscellaneous problems. The book is intensely interesting and will be of considerable value to both the specialist and the student.

G. Temple

Optimization Techniques with Applications to Aerospace Systems, *edited by George Leitmann. (Vol. 5 of* Mathematics in Science and Engineering). *Pp. xiii+453. Academic Press, New York and London. 1962. £5 14s. 6d. net.*

The impressive list of contributors to this work contains among others Richard Bellman and Robert Kalaba of the Rand Corporation, and John Breakwell of Lockheed Missiles and Space Co.

After a general survey of the theory of maxima and minima, the subsequent chapters discuss direct methods for obtaining optimum trajectories and for the extremization of linear integrals in flight mechanics by Green's theorem. Then the calculus of variations is applied to problems of aerodynamics and propulsion systems. Subsequent chapters discuss the methods of gradients, the application of the Pontryagin maximum principle, the optimization of trajectories by means of dynamic programming, and a survey of various computational considerations. The latter chapters discuss the general imbedding theory, impulsive transfer between elliptical orbits, and the optimum spacing of corrective thrusts in navigation techniques.

This work has been planned to appeal to the widest possible audience, including engineers, scientists, and applied mathematicians, and highly technological terminology has therefore been avoided. It should be of the greatest interest to all those who are concerned with the design of aerospace systems and also to those who are interested in the recent modern developments in the calculus of variations.

G. Temple

MATHEMATICAL TABLES

Handbook of Mathematical Tables (supplement to Handbook of Chemistry and Physics), *Editor-in-chief: C. D. Hodgman. Pp. x+579. Chemical Rubber Publishing Co. Cleveland. 1962. 70s. net.*

This almost unique volume contains all the usual mathematical tables covering such topics as logarithms, trigonometric functions, hyperbolic functions, reciprocals, powers, primes, and various statistical tables. But it goes much further. It is something of an encyclopaedia in that it defines and tabulates various special functions, Fourier series, integrals (both definite and indefinite), formulae of vector analysis, differential equations and their solutions, and various curves and surfaces. It will be a most valuable reference book for pure and applied mathematicians, as well as for physicists, chemists, and engineers.

The surprising feature of this book is the amount of material it contains. Many dictionaries and reference works on mathematics and allied subjects have appeared in recent years, but probably none matches this book in size and reasonableness of price.

L. S. Goddard

SOLAR PHYSICS

The Solar Corona, *edited by John W. Evans.* **(Proceedings of the International Astronomical Union Symposium No. 16, held at Cloudcroft, New Mexico, 1961.)** *Pp. xi+344. Academic Press, New York and London. 1963. £5 net.*

The conference reported here took place immediately after a period of maximum solar activity and was able to use much of the information gained during the I.G.Y. It resulted in an important discussion of our knowledge of the corona.

This volume is divided, as was the symposium, into three parts—local physics of the corona, coronal phenomena, and broader aspects of the corona. Each session began with a general survey, which was followed by original papers. These in turn were followed by discussions by study groups. it is unfortunate that these discussions are so briefly summarized. The symposium benefited from new results from radio-astronomy, but unfortunately only the early space-research results were available at the time for discussion of the studies of the ultra-violet and X-radiation of the corona.

The discussions leave an impression that coronal phenomena are more complex than had previously been suspected. Future work will be much helped by this volume.

A. Dollfus

POPULAR PHYSICS

Energy, *by Bruce Chalmers. Pp. xiii+289. Academic Press, New York and London. 1963. 46s. net.*

This book might be described as a reader's companion and refresher course for certain parts of popular scientific literature, chiefly American. Even recent developments such as fuel cells, magnetohydrodynamic generation of energy, fission nuclear reactions, and solar energy for spacecraft are included. So far as it goes, the book is clearly written and well illustrated, with its numerous diagrams kept deliberately elementary. Problems to be solved are provided at the end of every chapter. If these are taken seriously and if the text is read thoroughly, the reader will not remain technologically illiterate about many contemporary developments.

Primarily the book is intended as a background text for American undergraduate courses, to serve as an introduction to the physical sciences; its standard is equivalent to that of the fifth form in Britain. The wider significance of energy in a technological civilization is not dealt with at all adequately: readers are treated in this respect as though the United States of America formed no part of a modern tightly-knit world. Such naïve isolationism in a matter of such widespread importance is a pity. Were it not for this, the careful and elementary presentation of the text might make it useful for a variety of non-scientific readers.

A. R. Ubbelohde

MASS SPECTROMETRY

Advances in Mass Spectrometry (Proceedings of the Mass Spectrometry Conference, Oxford), *Vol. II, edited by R. M. Elliott. Pp. xvii+628. Pergamon Press Ltd, Oxford. 1963. £7 net.*

Mass spectroscopy has applications in many fields of science and technology, and regular conferences are valuable as a means of surveying the recent work, particularly in the absence of any regular publication suitable for reporting the whole range of studies. The conference held at Oxford in 1961, which is reported in this volume, was attended by almost 300 practising mass spectroscopists from all parts of the world. The organizing committee is to be congratulated, both for realizing that the advances of the three years since their previous conference called for an up to date survey of the field, and for the wide coverage and high standard achieved by their selection of speakers and topics.

Especially valuable are the papers in which authorities on various topics are given the opportunity to review their subject from its inception, so as to indicate how it has developed through the interplay of experiment and theory; to indicate new problems which have emerged; and to discuss the means of solving them. The reviewer found two of them: Field ionization mass spectrometry, by H. D. Beckey, and Current status of the statistical theory of mass spectra, by H. M. Rosenstock and M. Kraus, particularly stimulating.

This permanent record of the papers and the discussion that followed them will form an essential work of reference for all mass spectroscopists, although the specialist nature of the book makes it of limited interest to scientists in general. *J. H. Beynon*

CHEMICAL KINETICS

How Chemical Reactions Occur, an Introduction to Chemical Kinetics and Reaction Mechanisms, *by Edward L. King. Pp. xi+148. W. A. Benjamin Inc., New York. 1963. $2.15 paper, $4.35 cloth.*

This volume is one of about fifteen in a General Chemistry Monograph series, each written by someone not only well versed in the subject through his own research, but with experience in teaching it to college students. The volumes are well produced, with clear diagrams in two colours, and include numerous solved examples. This book is intended for first-year students with a background of high-school chemistry. It would be readily followed by students with G.C.E. A-level chemistry and would be useful in third-year sixth forms and during first-year university courses.

It deals first with the idea of reaction mechanisms, and goes on to describe the kinetics of a number of selected reactions. Having established the rate laws, it discusses theories of reaction kinetics and mechanisms. Reversible reactions and chemical equilibria receive a rather short treatment, and there are chapters on catalysis (including enzymes), very fast reactions, and the use of isotopes in the study of kinetics and reaction mechanisms. There is also a glossary and a reading list. *G. Van Praagh*

ANALYTICAL CHEMISTRY

Handbook of Analytical Chemistry, *edited by Louis Meites. Pp. xv+1806. McGraw-Hill Book Co. Inc., New York. 1963. £18 8s. net.*

This book attempts, so far as is possible within its compass, to cover the general field of analytical chemistry. There are fifteen sections in which the whole of the information is in the form of tables. They include information that is easy to tabulate, such as atomic weights and equilibrium constants and, at the other end of the scale, information, such as conditions for gravimetric precipitation, that is not so readily tabulated, but which nevertheless has been well compressed into this form.

It is inevitable that there should be some errors in a massive compilation of this kind, but these are remarkably few. Some of the figures for redox potentials are out of date; a few of the textbooks quoted in the bibliographies are of little use as works of reference, and some are out of date; and not all the more recent authoritative textbooks have been mentioned. Nevertheless, as a means of obtaining information quickly, if superficially, on most general matters of analytical chemistry, the book is good. It is a useful contribution to the literature of analytical chemistry. *R. Belcher*

Analytical Chemistry 1962 (The Proceedings of the International Symposium held at Birmingham University, April 1962), *edited by P. W. West, A. M. G. Macdonald and T. S. West. Pp. xi+411. Elsevier Publishing Co., Amsterdam. 1963. 90s. net.*

The papers are arranged under the following headings: spot tests and qualitative analysis (12 papers); semi-quantitative spot test methods (3 papers); organic reagents and their application (15 papers); electrochemical methods of analysis (9 papers); radiochemical methods of analysis (3 papers); optical methods of analysis (11 papers); separation processes (4 papers); titrimetric analysis (3 papers); and miscellaneous papers (3 papers). There are also a biographical sketch of Professor Feigl by H. A. Suter, and a lecture by R. Belcher on the applications and implications of specific, selective, and sensitive reactions.

As is only to be expected in any meeting as large as this one, the papers vary in quality and significance. The general impression, however, is that the level is appreciably higher than average, and the authors who contributed represent worthily the broad field of pure and applied analysis throughout the world. The quality of production is, as expected from these editors and publishers, very high.

Professor Feigl need have no qualms about accepting this volume as a fitting tribute to his work. *Cecil L. Wilson*

LANTHANIDE CHEMISTRY

The Chemistry of the Lanthanides, *by Therald Moeller. Pp. ix+117. Reinhold Publishing Corporation, New York; Chapman and Hall Ltd, London. 1963. 16s. net.*

It is arguable that the recent revival of interest in the lanthanides and actinides has not yet found adequate expression in books. In this little paperback, Professor Moeller goes some way towards redressing this situation, but his contribution is necessarily restricted by the limited space that he has allowed himself.

The book deals with the history, chemical properties, and isolation of the lanthanides, and it concludes with a short discussion of the actinides. It is pitched at the correct level for the honours degree student, but will also be read with profit by anyone who wants a good clear introduction to the subject. The treatment is up to date. For example, the traditional separation procedures, now of mainly historical interest, are not overemphasized at the expense of modern methods. However, the discussion of the magnetic properties and spectra of the lanthanides is too brief by the present-day standards of inorganic chemistry.

The low price of this work puts an adequate treatment of a special topic within the purchasing power of the average student.

F. A. Hart

ORGANIC CHEMISTRY

Topics in Organic Chemistry, *by Louis F. Fieser and Mary Fieser. Pp. xii+668. Reinhold Publishing Corporation, New York; Chapman and Hall Ltd, London. 1963. 80s. net.*

The evolution of the Fieser complex continues, and this latest attempt to keep the reader up to date, an increasingly hard task, is a kind of supplement to the authors' 'Advanced Organic Chemistry'.

It is in many ways, while conforming to the recognizable Fieser type, the most attractive of the whole series. Nine chapters in Part I deal rather fully with topics such as alkaloids, vitamins, and chemotherapy, which received less than their due in 'Advanced Organic Chemistry'. Part II is concerned with highlights among the recent developments such as ferrocenes, depsipeptides, tropone and related substances, and protection of carboxylic acids, to take a few entirely at random.

While Part I enables the student to round off his reading of the subject and achieve a broader coverage, Part II is adapted partly for this purpose, but perhaps still more as a mine that can be profitably worked for new research ideas. Among these some jewels will doubtless be found whose value and brightness may not be so clear when they are picked from the printed page, but which will shine with a new light when placed in the setting of the special circumstances of an investigation.

The relation between this book and 'Advanced Organic Chemistry' is a close one, and this is emphasized by the provision of back-references.

This book can be warmly recommended to advanced students of organic chemistry (to their teachers also) and to research workers and their supervisors at all stages.

R. Robinson

CARBOHYDRATE CHEMISTRY

Structural Carbohydrate Chemistry, *by E. G. V. Percival (second edition, revised by Elizabeth Percival). Pp. xv+360. J. Garnet Miller Ltd, London. 1962. 40s. net.*

The first edition of the book made its welcome appearance in 1950 and quickly found acceptance in chemical and biochemical disciplines as a general text on carbohydrate chemistry. Dr Elizabeth Percival has now produced an extensive revision of this textbook, incorporating many of the important and fundamental advances in carbohydrate chemistry that have appeared over the past decade. The book has been enlarged from 246 to 360 pages. It uses the same method of approach as in the first edition, but has a better arrangement of the chapters.

The first four chapters on monosaccharides have been modified a little in order to include conformational aspects, revised structures of osazones and their derivatives, and some new material on amino sugars. The subsequent chapters have been completely rewritten. The original chapter devoted to oligosaccharides has been made into two chapters, the first on disaccharides and the second on other oligosaccharides. The five chapters on polysaccharides give one of the best up to date accounts of this group of natural polymers. The liberal use of clear, well-printed formulae and diagrams is another good feature of this book, although I would prefer to see formulae used that avoid the use of bent covalencies, and that avoid, in particular, Fischer projection diagrams of the ring structures of monosaccharides with two such bends per covalency. There is an author index, an extensive subject index, and a useful selection of references quoted at the foot of each page.

The new 'Percival' will undoubtedly share the success of its predecessor, as it is, in my opinion, the best textbook on carbohydrate chemistry now available.

L. Hough

Methods in Carbohydrate Chemistry, *Vol. I,* **Analysis and Preparation of Sugars,** *edited by Roy L. Whistler and M. L. Wolfrom. Pp. xxii+589. 1962. £7 3s. net; Vol. II,* **Reactions of Carbohydrates,** *edited by Roy L. Whistler and M. L. Wolfrom. Pp. xv+572. 1963. £6 19s 6d. net; Vol. III,* **Cellulose,** *edited by Roy L. Whistler. Pp. xvi+407. 1963. £5 10s. 6d. net. Academic Press, New York and London.*

The editors of these volumes have set out to bring together, in a concise manner, 'recipes' for making a very wide variety of sugars and their derivatives. They have secured the collaboration of the world's leading carbohydrate chemists, who themselves have invented or developed the methods described. The volumes provide a wealth of information in a brilliant way.

Carbohydrates cannot stand rough treatment, so that special methods are needed to bring about, for example, crystallization, purification, reactions. In Volume I are described the methods of preparation and analysis of the simple monosaccharides. A preliminary section is provided on laboratory techniques especially applicable to carbohydrates, such as purification, crystallization, and gas-liquid partition chromatography. Important instructions are precisely given for paper chromatography and all other chromatographic methods.

It is these methods and the neat techniques of applying them that in recent years have made an enormous difference to the separation of closely related sugars and their derivatives. A brief description is provided on the preparation of certain labelled sugars, and it is rather to be regretted that this chapter has not been made more extensive.

The very complex but valuable method of applying periodic oxidation is described with great clarity, and the significance and value of colour reactions is demonstrated. Physical methods are provided in detail. Volume I is therefore of the greatest importance to practising chemists and biochemists, and particularly to students.

Volume II deals with the methods of preparation of derivatives of monosaccharides, and gives too some information on oligosaccharides. Sugar phosphates are dealt with in great detail. Short methods for the provision of some methyl ethers of mono- and disaccharides are concisely described, though unfortunately the valuable method of Kuhn has been omitted.

Volume III is essentially a new textbook on cellulose written by industrial chemists. Cellulose continues to be the world's most widely used fibre material, and vast industries depend on it. The isolation of pure cellulose from a variety of sources is concisely described and methods given for its analysis by chemical means and physical methods.

M. Stacey

STEROID CHEMISTRY

Steroid Reactions; An Outline for Organic Chemists, *edited by Carl Djerassi. Pp. vi+657. Holden-Day Inc., San Francisco. 1963. $9.75.*

The exponentially increasing amount of factual information in organic chemistry causes one to look with great interest at any novel method for systematizing this knowledge. This book consists of fourteen chapters, each containing a comprehensive summary of a group of steroid reactions. The information is presented in terms of structural formulae, along with a reference for each reaction. This presentation is ideal for rapid examination. The book contains some 2000 references, and literature coverage is virtually complete to the end of 1962.

Two chapters deserve special mention. The first provides a wide survey of the methods available for selectively blocking carbonyl and hydroxyl groups in polyfunctional molecules, and the second gives an extensive account of photolytic steroid transformations.

Many natural-product chemists will already know this work which contains such a wealth of information as to make the book of value to every organic chemist. All concerned in the production of this excellent book deserve to be congratulated on the speed with which it has appeared, and the book itself merits every success.

P. J. Sykes

VIROLOGY

Mechanisms of Virus Infection, *edited by Wilson Smith. Pp. ix+368. Academic Press, London and New York. 1963. 80s net.*

This book is a concise account of many important aspects of modern virology. It is likely to interest only virologists whose concern is the study of viruses in the laboratory, and it certainly cannot be read critically by anyone who is unfamiliar with the current literature on the subject.

Some maintain that all the main problems of virology have already been solved and that what remains is a trivial exercise. The study of this book does not support such a conclusion; on nearly every page there are speculations or debatable conclusions. The eight authors of the book are biologists, and so biochemical details are not laboured. Perhaps this is a good thing, because in recent years biologists have been subjected to a barrage of biochemical, biophysical, and statistical expertise. The biologist still has much to contribute and need not despair.

There is an excellent chapter on bacteriophages, and the inclusion of this chapter in a book otherwise wholly devoted to animal viruses is quite justified. Other chapters give an account of the pathways of virus infection, of mechanisms of cell infection, of virus pathogenicity, and of host resistance. This book is beautifully produced and illustrated, and there is an extensive bibliography which includes papers published in 1963.

F. Fulton

BIOCHEMISTRY

Mechanisms of Organic and Enzymic Reactions, *by S. G. Waley. Pp. xiv+365. Clarendon Press, Oxford. 1962. 70s. net.*

With the growing rapprochement between chemistry and biology, there is a need for books that will introduce biologists to the principles of organic chemistry, and organic chemists to those of biochemistry. Dr Waley has set out to meet this need in an original manner by writing a joint account of chemical and biochemical reaction mechanisms as an integrated whole.

The first two chapters of the book consist of a conventional introduction to valence theory, structural chemistry, and reaction kinetics. Each of the following chapters deals with one general type of organic reaction, presenting first the evidence for its mechanism and then giving examples of enzymic reactions that are thought to take place in this way. The specific chapter headings are: nucleophilic substitution; olefine-forming eliminations, and additions to olefinic double bonds; additions to aldehydes and ketones; esterification and hydrolysis; further reactions of acids and their derivatives; saturated rearrangements; aromatic substitutions; and mechanisms of some polymerizations. The book is well written, and the production is up to the usual high standards of the Clarendon Press.

It must be admitted straight away that this would not be a satisfactory text for a course on organic reaction mechanisms. Too many important topics are omitted, such as the chemistry of free radicals and carbenes, and the presentation of the rest is superficial and unoriginal. Much of the book gives the impression of having been compiled by a 'scissors and paste' technique from familiar sources. For the same reason, one could not recommend the book as an introduction to biochemistry for organic chemists; far better books of this kind are already available.

Nevertheless the book should serve well as a text for courses on biochemistry, and this, no doubt, was its primary purpose. The idea behind it is excellent, and the presentation very clear.

M. J. S. Dewar

PLANT PHENOLICS

Enzyme Chemistry of Phenolic Compounds (Proceedings of the Plant Phenolics Group Symposium, Liverpool), *edited by J. B. Pridham. Pp. ix+142. Pergamon Press Ltd, Oxford. 1963. 50s. net.*

The somewhat cryptic title covers thirteen papers given by various authors at a symposium of the Plant Phenolics Group in April 1962. The papers vary in clarity. Among other things, they deal with polyphenols as enzyme substrates, and conversely with phenolases as tools for the study of the organic chemistry of tannins. They show that the physiology of polyphenols in plants is still as vast and obscure a subject as ever and that the same is probably true in respect to microorganisms. References to animals are in the main excluded. The author of the paper on tocopherols estimates that this has reduced his field by 99·9 per cent. The tocopherols afford yet another example of substances produced by plants and of great significance in animal physiology (in this case as the fertility hormone), but whose importance in the simpler organization of plants is still elusive. Other subjects dealt with are the action of phenols as enzyme inhibitors, the special precautions necessary in the purification of phenolases, and the hydrozylating activities of the cresolases. There is a particularly fair and lucid account of ubiquinone and plastoquinone as putative electron transporters in, respectively, plant respiration and photo-synthesis. This is written by the Liverpool team working on them.

W. O. James

CARCINOGENESIS

Chemical Carcinogenesis, *by David B. Clayson. Pp. viii+467. J. & A. Churchill Ltd, London. 1962. 72s. net.*

The author is to be congratulated on so effectively presenting an account which, while surveying the subject in a comprehensive and critical manner, does not obscure fundamentals by excess of facts. The outstanding quality of the book is its objectivity in assessing conflicting and doubtful data. The book can be highly recommended as a guide and source of reference to those engaged in the field of chemical carcinogenesis, for whom chapters 3 and 4, on methodology, will be particularly helpful. It should also prove useful to investigators in other branches of cancer research, and to clinicians, to whom the excellent chapter on environmental cancer in man will have special interest. More than half the book is devoted to the carcinogenic properties and metabolism of various classes of compounds; this is followed by a detailed consideration of modifying influences in carcinogenesis and of a critical appraisal of the current theories on the mode of action of carcinogens. The chemotherapeutic action of some chemical carcinogens is also discussed.

It is hoped that in the next edition the brief introductory chapter on the nature of cancer will be revised and expanded, and that an additional chapter will be added on chemical carcinogenesis and viral action.

I. Berenblum

SOIL FUNGI

Soil Fungi and Soil Fertility, *by S. D. Garrett. Pp. viii+165. Pergamon Press, Oxford; The Macmillan Company, New York. 1963. 10s. 6d. net.*

When a world authority writes an introductory book about his own field of work, the result may be an outstanding contribution to the better understanding of his subject by a wide public. This book is a case in point. No one who knows Dr Garrett will be surprised at the excellent clarity of expression and the felicity of phrase with which he has explained the intricate inter-relationships existing between micro-organisms living in the soil. Many of the techniques involved have been developed by Dr Garrett himself and by those associated with him, and thus he writes with particular authority on the conclusions derived from experiments which have followed their exploitation. About one-third of this book is devoted to a consideration of the physiology, developmental morphology and breeding systems of fungi, which serves as an introduction to the intricacies and peculiarities of fungi vis-à-vis the 'higher' plants on the one hand, and animals on the other. The book assumes only a minimum of biological knowledge, and yet is an intriguing account for those mycologists who have an interest in the soil.

R. L. Lucas

CHEMICAL TAXONOMY

Chemotaxonomie der Pflanzen, *by R. Hegnauer Vol. I.* Thallophyten, Bryophyten, Pteridophyten und Gymnospermen. *Pp. 517. Birkhäuser Verlag, Basle and Stuttgart. 1962. SFr.96.*

The study of the chemical constituents of plants in relation to their taxonomic position is a subject that is now developing rapidly. Chemical taxonomy not only provides guidance to the organic chemist in his search for new natural products, but also assists the systematic botanist with the examination of taxonomic problems by processes that are complementary to the classical methods normally used.

This book is an exciting contribution to plant biochemistry. It summarizes the available information on the natural products isolated from representative Thallophytes, Bryophytes, Pteridophytes, and Gymnosperms. Almost half of the book is concerned with plants belonging to the first three groups, and the rest is devoted to the Gymnosperms. As would be expected with such a wide range of plant forms, this volume contains a tremendous amount of information about many types of natural product. This information is presented in a form that will certainly appeal to the natural-product chemist, but it is possible that some botanists may object to certain features of the taxonomic classifications that have been used. The book is not just a catalogue of information, but discusses the biosynthesis of various types of natural products, as well as giving a very adequate summary of reviews of phytochemistry, systematic botany, and plant anatomy.

W. D. Ollis

BIOMETRY

The Essence of Biometry, *by John Stanley. Pp. xiii+147. McGill University Press, Montreal. 1963. $5.*

Professor Stanley has produced an account of elementary statistical techniques directed to the graduate student in zoology. The exposition is verbal rather than algebraic, and extends to the one-way analysis of variance, so that factorial experimentation and experimental design are not covered. However, the pleasantly informal exposition may help to tide the reader over the difficulties. Some topics not often included in elementary works are treated in detail, notably curve-fitting; more space is devoted to the fitting of an untransformed sigmoid than to the analysis of variance. Transformations, which are in regular use to obviate curve-fitting, are little discussed: the probit transformation is the only one dealt with in detail.

Some of the book's terminology and notation differ from that of similar and more advanced presentations, and may make further reading by the student more difficult by preventing him from recognizing the conventional terms when he meets them elsewhere.

Professor Stanley's approach to biometry may appeal to zoologists who find the standard texts too difficult; the book would benefit from revision, from a closer organization, and from modernization; it reflects the essence of biometry as it was two or three decades ago rather than as it is now. Printing, paper, and proof-reading are excellent and the index is adequate; but the bibliography is limited and inappropriate to the text.

R. E. Blackith

ENDOCRINOLOGY

An Introduction to General and Comparative Endocrinology, ***by E. J. W. Barrington. Pp. xiii+387. Clarendon Press; Oxford University Press, London. 1963. 35s. net.***

General endocrinology is designated as dealing with the principles that have been derived from the study of man and other eutherian mammals. The term comparative endocrinology is a recent one, though the initiation of the field of study in a modern context may be regarded as having stemmed from Berthold's (1849) observations on the capon. Barrington's book is the third of a triad appearing over the past year or so and specifically concerned with the endocrinology of non-mammalian vertebrates and invertebrates.

In its fifteen chapters the book covers all the major topics. Though there are necessarily omissions in detail, this volume will serve as an excellent introduction to a subject that is commanding more and more attention. Great care has clearly been taken to give a lucid and logical development of the subject. A high standard is reached, and those parts concerned with the thyroid and the islets of Langerhans, being among the research interests of the author, are particularly authoritative. The figures and line drawings are helpfully conceived and add considerably to the exposition of the text. Some of the half-tone plates are less successful. Wisely the cost of the book has been kept low, by modern standards, so that on the counts of both content and price it can be warmly recommended for student use. *I. Chester Jones*

MARINE MICROBIOLOGY

Marine Microbiology (Deep Sea), ***by A. E. Kriss (translated from the Russian by J. M. Shewan and Z. Kabata). Pp. xviii+536. Oliver and Boyd, Edinburgh and London. 1963. £6 6s. net.***

This translation of Professor Kriss's monumental work will bring a great deal of new information to English-speaking bacteriologists, and it is therefore a welcome addition to the sparse literature on marine microbiology. Various chapters deal with the morphology, taxonomy, reproduction rate, biomass, and distribution of marine bacteria, while others are concerned with the effect of these bacteria on the chemistry and biological productivity of the sea. One chapter is devoted to the author's hypothesis that the variation in the number of bacteria within a water mass can indicate the origin of the water and the course of oceanic currents.

However, the book fails to be convincing, not because of what it says but because of what is left unsaid. Frequently the reader is presented with unsubstantiated statements. Thus the results of many bacterial plate counts are recorded, but no reference is made to the statistical limitations of this method. There are pages of diagrams of various bizarre organisms which would have been more satisfactorily displayed as photographs. Many of the conclusions are apparently based on single observations. In short, this book can be regarded as one indicating the scope of marine microbiology rather than as an authoritative work. *G. D. Floodgate*

MARINE BIOLOGY

Advances in Marine Biology, ***Vol. I, edited by F. S. Russell. Pp. xii+410. Academic Press, London and New York. 1963. 84s. net.***

The contents of this first volume illustrate the editorial policy of including reviews of recent advances in fisheries biology and of research on marine organisms generally.

V. L. Loosanoff and H. C. Davis discuss conditioning bivalves, the rearing of bivalve larvae, and factors affecting development. The techniques developed raise exciting possibilities for investigating genetic strains.

Although the recent controversy over the European and American freshwater eel is not settled by the paper by the late Anton Bruun, the re-examination of several factors appears to favour granting specific rank to the two geographical populations.

Reviewing photo-reception in fishes, J. A. C. Nicol discusses retinomotor movements and the variety of rod pigments. Aspects of behaviour discussed range from colour change in fishes to vision as a factor in fish schooling. C. M. Yonge's paper emphasizes the diverse biological implications of coral reef study. Variation in growth forms of coral, and the significance of zooxanthellae, especially in relation to growth rate, are discussed. The contribution by J. H. S. Baxter and F. G. T. Holliday includes observations on salinity tolerance, osmo-regulation, and biochemical composition in clupeids, and examines the importance of vision in behaviour.

The volume is well produced and includes excellent bibliographies. The present reviewer considers, however, that even major papers might be limited to some 60 pages.

J. E. G. Raymont

INSECT PHYSIOLOGY

Advances in Insect Physiology, ***Vol. I, edited by J. W. L. Beament, J. E. Treherne, and V. B. Wigglesworth. Pp. xiii+512. Academic Press, London and New York. 1963. £5 5s. net.***

For many years Professor Wigglesworth and his colleagues have made elegant use of insect material for the study of cellular and comparative physiology. The appearance of this volume under the aegis of the Cambridge group is therefore fitting as well as timely.

Contributions review and critically discuss the acetylcholinesterase and possible associated systems in insect nervous function (E. H. Colhoun); nutrition and feeding of orthoptera (R. H. Dadd); biochemistry of the fat body (B. A. Kilby); electrophysiological studies of normal and poisoned central nerve tissue (T. Narahashi); chemical structure of the insect cuticle (K. M. Rudall); osmotic and ionic regulation (J. Shaw and R. H. Stobbart); and the structure and metabolism of the central nervous and neuromuscular tissue (D. S. Smith and J. E. Treherne). All the contributions reflect the authority and insight peculiar to the personally-involved research worker of some years standing. One can mildly criticize. Colhoun, for example, tends to ask questions instead of answer them. Again, in the discussion of permeability of the neural lamella by Smith and Treherne reference to the important observations by O'Brien in relation to the penetration of cholinesterase inhibitors was conspicuously absent.

The editors and contributors are to be congratulated on the excellence of the scientific content and its presentation. This volume and its successors are likely to appear on the shelves of all serious students of invertebrate biochemistry and physiology.

F. P. W. Winteringham

FISH ECOLOGY

The Ecology of Fishes, ***by G. V. Nikolsky (translated from the Russian by L. Birkett). Pp. xv+352. Academic Press, London and New York. 1963. 75s. net.***

The fisheries scientist seldom has enough information, particularly when faced with problems concerning the present and likely state of fish stocks and the reactions of fish, both to the inorganic and the living factors of their environment and to those factors introduced by man. Professor Nikolsky's book is written with just such points in mind, even to the extent of frequently selecting as examples fish of commercial value. Here, however, two considerations arise. First, the author visualized this book as complementary to his earlier 'Special Ichthyology', which dealt with the ecology of individual species, so that the present book sometimes gives a sketchy impression, and principles may seem to be enunciated on slender evidence. Secondly, the content of both is of particular relevance to the Soviet Union, as also is the present bibliography.

Only about 80 out of well over 300 references are non-Russian, and not all are those most to be expected. This is reasonable enough, as the book was designed for Russian readers, and Western books have insufficient Russian references. But many who will find this book of real interest and value would have benefited from a wider field of reference, particularly in vital sections, such as that on population dynamics, which has scarcely a reference to the considerable British and North American work in this field.

C. E. Lucas

PHARMACOLOGY

Progress in Drug Research, ***Vol. 5, edited by Ernst Jucker. Pp. 654. Birkhäuser Verlag, Basle and Stuttgart. 1963. SFr. 124***

Probably too many 'reviews' are written today, and many are too brief to be worth while and are duplicated by other publishers. But there will always be a need for articles such as those in this volume. This applies especially to the section on medicinal phenothiazines, which is an all-embracing account occupying half the space in this volume and quoting 6800 references. It deals with all aspects of the phenothiazines used in medicine including synthesis, chemistry, and metabolism as well as mode of action and application. In 102 tables all the examples of this class of compound known up to 1963 are listed.

The shorter articles (144 and 114 pages) deal with the anti-inflammatory action of hydrocortisone and with the relation between analgesia and addiction. Much information is presented on relations between structural changes and drug effectiveness, in addition to a discussion of possible modes of action. Again much of the data is in tabular form.

All the articles illustrate very well the current situation in drug research, which necessitates the addition of a synthesis of organic chemistry, physical chemistry, and biochemistry to the older pharmacological approach.

J. N. Smith

COMPARATIVE PATHOLOGY

An Introduction to Comparative Pathology, a Consideration of some Reactions of Human and Animal Tissues to Injurious Agents, ***by G. A. Gresham and A. R. Jennings. Pp. xi+412. Academic Press, London and New York. 1962. 84s. net.***

The authors have packed an enormous amount of information into this relatively small book. It follows that most of it is written in a terse and concise style. The effect of this is a highly concentrated piece of reading, which may be rather tough going for beginners, but will be stimulating for those who already have some background knowledge of human or animal pathology. It is difficult to assess, without further experience of the volume, how useful it will prove as a work of reference.

There are twelve chapters, arranged on a more or less aetiological basis. The first chapter is concerned with genetically determined disease, the next five with parasitic diseases, beginning with those caused by viruses and ending with those caused by helminths. There follow chapters on disease caused by inanimate agents (gases, dusts, chemicals, physical agents, and radiation), nutritional disease, the conditions associated with antigen-antibody reactions, and the effects of endocrine dysfunction and of ageing. The final chapter is a brief but well-selected review of the comparative pathology of neoplastic disease. It is not possible in a short review to go into detail about the merits of this book, but I feel that I will be returning to it periodically for some time to come.

J. W. Orr

APPLIED ELECTRONICS

Electronic Devices and Networks, ***edited by E. E. Zepler and S. W. Punnett. Pp. xi+217. Blackie & Son Ltd, London and Glasgow. 1963. 35s. net***

The books in the 'Electronic User Series' are aimed at those 'qualified scientists and engineers for whom electronics is a fringe activity'. A study of this volume in the series shows that the books are intended to provide an understanding of the detailed working of electronic circuits and systems and to enable the reader to design such circuits. Therefore the reader must be willing to work hard, and the books in general, and certainly this one in particular, are not for those who merely wish to learn enough to organize electronic engineers.

The eight chapters in this book have been written by six authors, and the general editors have not imposed uniformity on their colleagues. The descriptive chapters contrast markedly with those involving detailed mathematical argument; there are no examples for the reader to try, and the most difficult and compressed chapter—on electrical networks—has no examples in the text. It is doubtful if any reader will derive much benefit from this chapter, nor indeed from some of the other analytical sections, unless he has expert assistance. The book may be useful to someone attending a lecture course designed for users of electronic equipment; it is unsuitable for those studying privately.

J. T. Allanson

SPACE RESEARCH

A Review of Space Research. (The Report of the Summer Study conducted under the auspices of the Space Science Board of the National Academy of Sciences at The State University of Iowa, June–August 1962.) ***Pp. 577. Publication No. 1079, National Academy of Sciences – National Research Council, Washington, D.C. 1962. $4.***

Since 1945, when the first relatively crude upper atmosphere experiments were launched in captured V-2 rockets, space research in America has become the vast undertaking that is now identified by the general public mainly with manned space flight and communications satellites. Within this overall effort there exists also a programme of basic scientific research in space, so extensive in its present scope and future objectives that few scientists working directly in this field can hope to assimilate the total programme in detail. In 1961, the Space Science Board of the National Academy of Sciences and the National Aeronautics and Space Administration recognized the necessity of a comprehensive review of America's space research programme. Consequently, more than 100 scientists, both full- and part-time, attended a two-month summer study at the University of Iowa in 1962 to discuss critically the nation's space science activities. The publication contains a report of their discussions, findings, and recommendations. Its emphasis is on basic scientific research rather than on related technology, and in addition, it examines the financial and social implications of the American space programme. Manned space flight is discussed in terms of its scientific usefulness, and the study group, while remaining somewhat critical of the huge sums being allocated for this work, generally recognizes the scientific benefits that could result, particularly if the astronauts themselves are trained scientists. The report is a valuable work of reference to the specialist. The general reader will find its treatment of the underlying philosophy of American space research as interesting as its scientific content.

P. J. Bowen

HISTORY OF SCIENCE

Fourcroy: Chemist and Revolutionary, ***by W. A. Smeaton. Pp. xxi+288. Printed for the author by W. Heffer & Sons Ltd, Cambridge. 1962. 40s. net.***

Antoine-François de Fourcroy was an important chemist, a distinguished teacher, and a successful Civil Servant. His work was important to Lavoisier's 'new chemistry' in two ways—his research supported the new theory, and the clarity with which he expounded it was an important element in its success. He was ambitious, and the rôle he played during the revolutionary period and during the subsequent Empire gained him both fame and fortune.

W. A. Smeaton is already well known for his work on the history both of science and of teaching during the French Revolution, and his new book is an indispensable contribution to the history of science. The first part deals carefully and precisely with the life and times of Fourcroy, the second with his chemical studies on, for example, the inflammable gases and the composition of water. This part also covers the works—particularly *Elémens d'histoire naturelle et de chimie*, *Philosophie chimique*, and *Système des connaissances chimiques*—and the journals that he edited. This important book ends with an exhaustive and extremely useful bibliography.

Lucien Scheler

89

ENDEAVOUR

VOLUME XXIII MAY 1964

A review of the progress of science, published in five languages

Editor TREVOR I. WILLIAMS, M.A., B.Sc., D.Phil., F.R.I.C.
Foreign Editor G. N. J. BECK, B.Sc., A.M.I.E.E.
Assistant Editor W. A. OSMAN, B.Sc., M.Sc., A.R.C.S.

Contents

92:546.02
54 Editorial: Frederick Soddy and the concept of isotopes

546.118
55 Plasma chemistry
C. P. Beguin, Ing. Chem., Ph.D., A. S. Kanaan, B.Sc., M.Sc., Ph.D., and J. L. Margrave, B.S., Ph.D.

576.314:578.6
61 Electron microscopy of the cell surface
E. G. Gray, Ph.D.

551.345:624.139
66 Permafrost and related engineering problems
R. J. E. Brown, M.A., Ph.D. and G. H. Johnston, B.Sc.

538.6
73 Magnetohydrodynamics
W. B. Thompson, M.A., Ph.D.

551.579
81 The movement of soil water
E. C. Childs, B.Sc., Ph.D., D.Sc.

581.332:582.28
85 Germination of fungus spores
David Gottlieb, B.S., M.S., Ph.D.

537.533:621.385.833
90 Field-ion microscopy
D. G. Brandon, M.A., Ph.D.

615.783:547.94
97 The relief of pain—the search for the ideal analgesic
K. W. Bentley, M.A., B.Sc., D.Phil.

551.324
102 Glaciology
G. de Q. Robin, M.A., Ph.D.

107 Book reviews

Some books received (cover iii)

Communications to the Editor should be addressed to South Block, Thames House, Millbank, London S.W.1. Telephone: Tate Gallery 2181

Frederick Soddy and the concept of isotopes

The word isotope is now in such common usage that it is hard to realize that only fifty years have elapsed since it was first coined. In a letter to *Nature*, published on 4th December 1913, Frederick Soddy—then lecturer in physical chemistry in the University of Glasgow—advanced the idea that atoms could be chemically identical and yet have different weights; for such elements he proposed the name isotope, because they occupy the 'same place' in the Periodic Table. Because of the strict classical derivation of the word, from the Greek *isos topos*, it has passed virtually unchanged into every language of the world. According to Lord Fleck, who at the time was working as a young assistant in Soddy's laboratory, we owe its derivation to Dr Margaret Todd, a friend of the Soddy family. Soddy was expounding his ideas at an informal dinner party in the home of his father-in-law Sir George Beilby, the distinguished industrial chemist. In the course of conversation he sought a word to express his new concept, and Dr Todd immediately supplied it.

Although of no great scientific distinction, Margaret Todd deserves more than passing mention. Born in 1859, she entered medicine at a time when the profession's prejudice against women was still very strong. After training at Glasgow, Edinburgh, and Berlin she became a schoolteacher and then entered the Edinburgh School of Medicine for Women. This was founded in 1886 by Sophia Jex-Blake, a formidable protagonist of women's right to a medical career. She graduated at Edinburgh in 1894, and took also the M.D. of Brussels University. She became assistant physician to the Edinburgh Hospital for Women and Children, and also achieved considerable success as a writer of romantic novels, some with a medical theme, under the pen-name Graham Travers.

To mark the fiftieth anniversary of the first appearance of the word isotope in print the Chemical Society held a special meeting in Glasgow.[1] Very appropriately, personal tribute was paid to Soddy, who is to many a somewhat shadowy figure on the scientific stage. That he was both a genius and a master of experimental technique none can now doubt; equally, none can doubt that he was a strange man whose own impetuous actions created most of the difficulties that beset his life. In his later years he was embittered by the belief that the world had failed to accord him due recognition for his contributions to the study of radioactivity. Looking back, it is hard to believe that he was greatly wronged, for in 1921 he achieved the high honour of a Nobel Prize, specifically awarded for his contributions to the chemistry of radioactive substances and his investigations into the origin and nature of isotopes. At home, the Royal Society had elected him to its Fellowship in 1910 at the early age of thirty-three, and his own university awarded him an honorary degree. Although his later years were ones of great activity, his time was increasingly devoted to developing social, political, and economic theories that gained no general acceptance.

Without in the least detracting from the greatness of Soddy's achievement, it is fair to say that the concept of isotopes was, as it were, ready to be snatched from the air by a man brilliant enough to see the significance of new discoveries and brave enough to stake his professional reputation in flying in the face of ideas so long established that they tended to be widely looked upon as eternal truths. As long ago as 1886, indeed, William Crookes pointed out that the existence of atoms of the same chemical properties but different masses would explain the known departures from Prout's law. However, it was in the context of radioactivity that the concept became of renewed interest some twenty-five years after Crookes's prophetic inspiration. Earlier in the same year as Soddy published his letter to *Nature*, the Polish chemist K. Fajans published in the *Physikalische Zeitschrift* a paper in which he accurately formulated the Group Displacement Law, but wrongly associated the mechanism with the valency electrons. In Britain, the displacement law was independently enunciated by A. S. Russell, also in 1913. The first part of this law, that emission of an α-particle moves an element back two places in the Periodic Table, was contained in Soddy's 'Chemistry of the Radio-elements' of 1911. In July 1913, A. van der Broek put forward the idea that nuclear charge is proportional not to atomic weight but to the ordinal number of an element in the Periodic Table. He firmly concluded that an element's nuclear charge is determined by its place in the Periodic Table and that it is not a purely positive charge, but the difference between a positive and a smaller negative charge. Soddy's own letter acknowledged and developed van der Broek's ideas. H. G. J. Moseley, whose early death at Gallipoli was one of the tragedies of science, established beyond doubt that the place of an element in the Periodic Table is wholly determined by the nuclear charge. As Soddy put it: 'Moseley, as it were, called the roll of the elements.'

Soddy reached his scientific peak in 1913, in his middle forties, with his formulation of the concept of isotopes, but he had already carried out distinguished work. As a young man, he had been associated with many of the great school of British scientists who at the beginning of this century were rapidly developing great new fields of physics and chemistry. From 1900 to 1902 he worked with Rutherford at McGill University and with him published a series of exceedingly important papers on radioactivity. Together they reached the then revolutionary idea that 'radioactivity is at once an atomic phenomenon and the accompaniment of a chemical change in which new kinds of matter are produced'. By 1902, they were investigating the gaseous emanation of radium and suspected that it had the properties of an inert gas, a field of chemistry in which Sir William Ramsay was then pre-eminent. Anxious to pursue this line, Soddy sought, and received, Ramsay's permission to work with him at University College London.

Arrived in London, Soddy's immediate problem was that only the Curies in Paris had any pure radium. By chance, walking through a London street near the laboratory, he saw pure radium bromide offered at seven shillings and sixpence a milligram by a German firm. 'Even I could afford some at this price', he recorded, 'and I at once ordered 20 milligrams.' With this 20 mg, and a separate batch of 32 mg from the same source lent by Rutherford, a classic experiment was carried out on 8th July 1903. Spectroscopic examination of gas produced by the radioactive decay of the radium salt showed that it contained helium, the element whose existence had been assumed by Lockyer, thirty years earlier, to account for an unknown line discovered by him in the spectrum of the Sun's chromosphere.

This marked the end of the brief but fruitful collaboration between Ramsay and Soddy. In the spring of 1904 the latter went to Australia, returning in the autumn to take up his appointment at Glasgow. Ten years later he was elected to the chair of chemistry at Aberdeen, but his plans for research there were frustrated by the outbreak of war. In 1919, the wheel turned full circle; he returned, as Dr. Lees Professor of Chemistry, to Oxford, where in 1898 he had headed the honours list, with Ramsay as his external examiner. Hopes that he would establish a British school of radiochemistry there were disappointed. He became increasingly immersed in his new financial and political doctrines of social credit and retired from the chair in 1936, after his wife's death. He himself died in 1956, in his eightieth year.

---

[1] *Proceedings of the Chemical Society*, November 1963, pp. 325–31

Plasma chemistry

C. P. Beguin, A. S. Kanaan, and J. L. Margrave

It has been clear for a long while that many interesting and potentially valuable reactions will take place at extremely high temperatures, but available flames and refractories have normally imposed a limit of about 5000°C. Temperatures as high as 50 000°C can be obtained by plasma torches. This article discusses the design of such torches, the reactions they have made practicable, and their possible future industrial applications in a number of fields.

Within recent years, new devices—plasma torches—that use high-intensity electric arcs have been developed to produce gas streams at temperatures between 5000 and 50 000°C. The gases, which are at least partially ionized, are called plasmas. The 'chemistry' of such gases is an intriguing modern field that challenges the techniques available to both theorists and experimentalists.

Before these torches were developed, the highest temperatures which could be continuously sustained lay somewhere between 5000 and 6000°C, and indeed only a few chemical reactions could be performed above 4000°C. The principal heating devices for chemical processes were electric furnaces and flames, and their useful temperatures were limited, as the best available refractory materials have melting points hardly exceeding 4000°C. In certain flames, such as H_2+F_2, $(CN)_2+O_2$, and $(CN)_2C_2+O_2$, temperatures above 4000°C can be obtained but these materials are neither cheap nor easy to handle and are therefore not especially suitable for industrial processes. More usual flames, such as $C_2H_2+O_2$ and C_3H_8 or $C_4H_{10}+O_2$, yield temperatures in the 2000–3500°C range.

Low-intensity electric arcs have been used for a long time in chemistry and metallurgy, but the effective temperature in an open-arc channel does not exceed 6000°C, and only two homogeneous gas reactions in arc channels were used industrially before the introduction of plasma torches. These were the Hüls process for acetylene production and the Birkeland-Eyde process for nitrogen fixation.

Plasma generators

A plasma is an electrically neutral, partially ionized gas with properties that depend on pressure, temperature, and composition. Plasmas at pressures close to 1 atmosphere consist of molecules, atoms, positive ions, negative ions, and electrons; there is a chemical equilibrium between constituents, so that the composition is exactly defined for a given temperature and pressure. As the dissociation of molecules and the excitation and ionization of atoms are all endothermic processes, an increase in temperature will reduce the number of molecules present and increase the degree of ionization. Since the partially ionized gases conduct electricity, electric arcs are a convenient source of high temperatures. Several different types of arc device are described in the following sections.

The low-intensity arc and the Beck arc. The low-intensity arc was first operated around the beginning of the nineteenth century and its properties are well known. Three different regions can be distinguished in such an arc (figure 1). The cathode and anode fall-spaces are short regions, a fraction of a millimetre in length, and are characterized by a high potential-gradient. The fall-spaces are caused by the space-charge accumulation and the necessity for the transfer of charges across the junction between the plasma gas and the solid conductor. The column is the longest region, and the potential drop across it is approximately 70 per cent of the total potential difference, although the potential gradient is smaller than near the electrodes. As the three regions are in series, the same current flows through them and the power dissipated per unit volume is largest in the fall-space regions, where the potential gradient is highest. The energy input to the arc is dissipated by radiation and convection, and by conduction through the electrodes. At normal atmospheric pressure, the temperature of the column is about 5000–6000°C.

An increase in current density gives an increase in temperature. However, simply increasing the current does not necessarily increase the current density, because the volume of the arc and, in particular, the area of the anode spot (figure 2) also increase. If unrestricted, the arc will spread on the cylindrical sides of the anode after it has covered the planar anode surface. It can be prevented from spreading either by properly adjusting the arc gap or by protecting the sides of the anode by coating it with an insulator, by cooling it with a gas stream, or by putting it inside a cooled copper ring. In the restricted arc, an increase in the current density results in an increase in the anode temperature.

When the material of the anode is heated to its boiling point, an entirely different form of arc, the Beck arc, appears. A long stream of vapour arises from the anode and gets longer

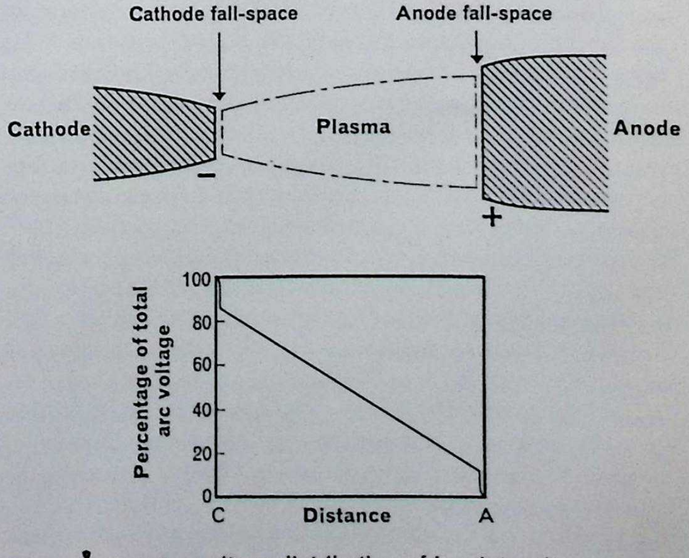


Figure 1 Relative voltage distribution of low-intensity arc.

C. P. Beguin, *Ing. Chem., Ph.D.*
Was born in 1929 and educated at the Federal Institute of Technology (Zurich). He joined the rare metals department of Ciba (Basle) in 1960 and moved to Rice University in 1963 as a Post-doctoral Research Fellow.

A. S. Kanaan, *B.Sc., M.Sc., Ph.D.*
Was born in 1931 and educated at Baghdad University, Colorado State University, and the University of Wisconsin. He moved to Rice University in 1963 as a Post-doctoral Research Fellow.

J. L. Margrave, *B.S., Ph.D.*
Was born in 1924 and educated at the University of Kansas (Lawrence). He was at the University of California (Berkeley), 1951–52, and the University of Wisconsin (Madison), 1952–63, before joining the staff of Rice University where he is Professor of Chemistry, and heads a research group in high-temperature chemistry.

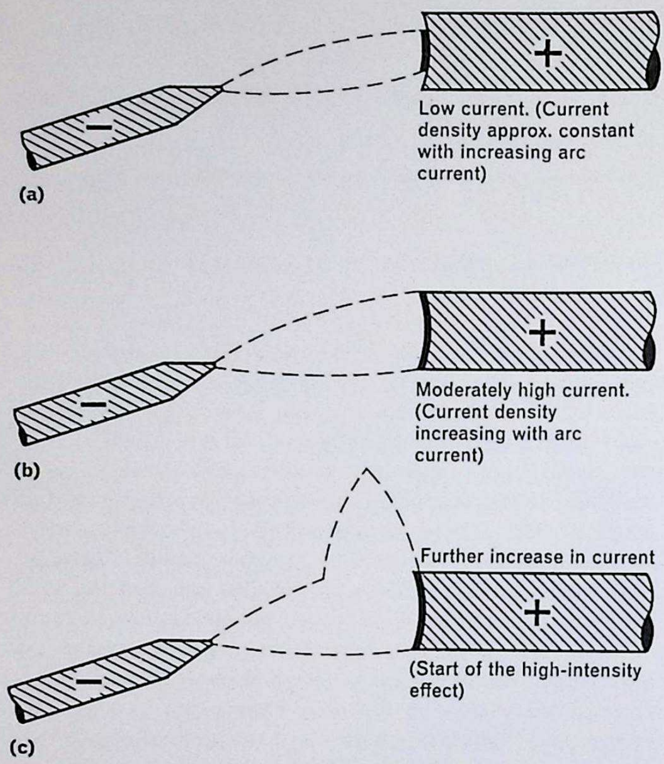

+
−
Low current. (Current density approx. constant with increasing arc current)
(a)
+
−
Moderately high current. (Current density increasing with arc current)
(b)
Further increase in current
+
−
(Start of the high-intensity effect)
(c)

Figure 2 Appearance of a carbon arc as the total current is increased through transition to the high-intensity arc. (*a*) Low-intensity region. (*b*) Just below transition point. (*c*) Just above transition point.

and brighter as the current increases. Lengths of several feet, temperatures of about 7000°C, and velocities of 5000 cm/sec have been recorded. For the low-intensity arc, the arc voltage decreases with increasing current, but this trend is reversed in the Beck arc, and 70 to 90 per cent of the total energy input may be dissipated on the anode. This energy is too high to be dissipated only by radiation and by conduction through the anode, and the excess is taken away by the jet of hot gases.

This high-intensity arc, discovered by Beck in 1910, has been used as a light source. It has also been utilized as a metallurgical tool where reactants are fed into a composite anode. However, applications in the chemical fields are somewhat limited by the unavoidable need for carbon (or some other conductor) in the anode to carry the current.

The Gerdien arc and the d.c. plasma torches. H. Gerdien and A. Lotz discovered a new kind of discharge in the course of experiments designed to improve the high-density arc. This was a fundamental step in the development of plasma devices. They started their experiments with an arc whose anode was inserted in a copper ring, the inner wall of which was cooled by water. They moved the ring along the arc column, and observed a phenomenon that was completely independent of the electrodes. In the ring, the arc channel was constricted because its periphery was cooled by the water: this is called a thermal pinch. The current density at this point increased to 100 A/mm² and the temperature was the highest obtained at that time. The arc burned in water vapour, and the bright, hot gases—plasma—emerged from both sides of the ring as jets.

This type of discharge was, apparently, never applied in chemical processes. However, its principle—the increase of current density and of temperature by thermal pinch—is the basis of the modern plasma jets.

If a cylindrical arc-channel is limited by a boundary wall (a cooled metal ring, a liquid, or a gas vortex flow) at a fixed temperature, it is completely characterized by the current flowing through it, the temperature, and the radius of the boundary cylinder. Above a given current, the field strength increases with the current if the radius is small. The smaller the radius, the more pronounced is this increase. Either an increase in current for a given radius, or a decrease of radius for a given current, corresponds to an increase in current density, power input, and axial temperature. Thus, somewhat paradoxically, by cooling the periphery of an arc channel for a given current, one increases its axial temperature. In addition to this thermal pinch, there is a magnetic pinch at high current densities, because the arc channel is constricted further by the self-induced magnetic field.

The modern plasma jet [1] is a modification of the Gerdien and Lotz arc. It is approximately cylindrical in shape (figure 3). One electrode is in the form of a conductive plate with a central hole, at one end of the cylinder. The other is a conductive rod that lies along the cylindrical axis and is held in position by an insulating plate at the other end of the cylinder. A gas is injected tangentially and forms a vortex in the chamber. When the arc burns, the hot gases emerge through the orifice and take away any energy that has not already been lost to the water-cooled walls of the torch. To minimize this energy loss, the hot gases have to emerge at a very high speed through the pierced electrode. For a given power input and gas flow rate, the diameter of the orifice determines the velocity, which is generally fairly high, of the jet. The d.c. plasma torches manufactured at present produce gas streams with temperatures between 8000 and 50 000°C and velocities between subsonic and Mach 20, at power inputs between 0·5 and 3000 kW. They are being used for welding, cutting, spheroidizing, and spraying metals, for the simulation of re-entry conditions in space research, and for chemical syntheses.

The radio-frequency plasma torch. In the r.f. plasma torch [2] a stream of ionized gas contained in a circular quartz pipe is heated by induction, a process essentially analogous to the induction heating of a metallic cylinder (figure 4).

As gases are non-conductors at low temperatures, the discharge must be initiated by the introduction of charge-carriers. There are several ways of doing this, the most usual being to heat inductively a graphite rod placed along the axis of the induction coil. When the rate of emission of thermal electrons is sufficient, a pilot discharge appears, and this can couple with the r.f. field to produce a more powerful discharge. If the graphite rod has a pointed tip, the pilot discharge may be a corona discharge. Once the main discharge is established, the graphite rod can be removed. The gas stream carries the ionized gas away from the induction coil in the form of a plasma flame. A fraction of the ionized gas must be recirculated, or the discharge will cease, and this recirculation is

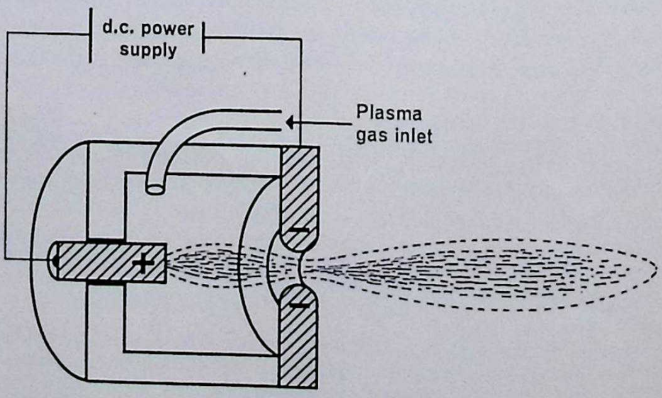


Figure 3 Direct-current plasma torch.

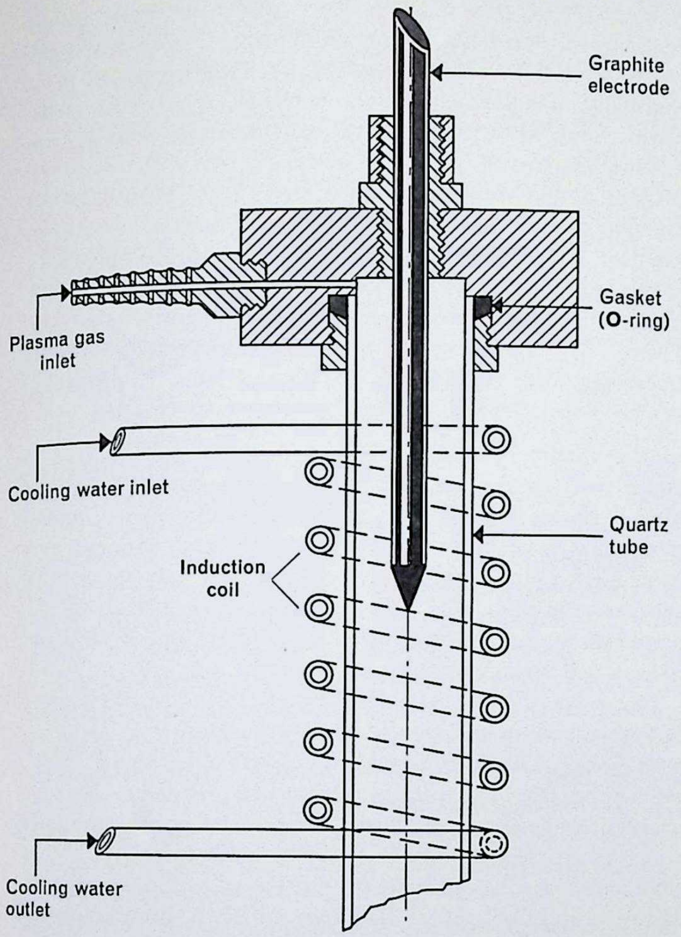


Figure 4 Radio-frequency plasma torch. The quartz tube is of approximately 22 mm bore.

achieved by introducing the gas tangentially, so as to create a low-pressure zone on the axis of the pipe. This forces some of the plasma to flow up the tube in the opposite direction to the main flow and stabilizes the discharge. When stabilized, the system works like a transformer, with the induction coil as a multi-turn primary and the plasma as a one-turn secondary.

The choice of the radius of the tube is governed by a peculiarity of r.f. induction heating: the induced current has a tendency to crowd to the outer surface of the load. There is a thin layer called the skin in which 86 per cent of the energy is dissipated. The efficiency of heating of a cylindrical load of radius r drops very rapidly when $r/\delta < 1{\cdot}75$, where δ is the skin thickness. On the other hand, when $r \gg \delta$, most of the energy is dissipated at the periphery and the bulk of the gas is not heated. At 5 Mc/s, the skin thickness for argon is about 2·5 mm and a tube with a diameter of 25 mm can be successfully used. The cold wall of the confining tube cools the external layers of the plasma which then has a tendency to contract. The higher the thermal conductivity of the gas, the more pronounced is this effect.

Since there are no electrodes in the r.f. plasma torch, it is possible to work with reactive gases, such as oxygen and air, that would erode the electrodes of a d.c. plasma torch at high temperatures. The only cooling necessary is on the outside wall. The absence of electrodes makes it easy to introduce solid substances along the axis of the plasma, and this is especially important for the Verneuil process for growing single crystals of metals and oxides. In addition, the gas velocity is usually small, allowing longer contact times between reacting gases than in d.c. plasma torches.

At the moment the most powerful r.f. torch is heated with a 200 kW induction heater and frequencies in the 1–25 Mc/s range have been used. Axial temperatures of about 16 000°C have been reported for an argon r.f. plasma.

Successful syntheses with plasma devices

Current research in the arc synthesis of compounds largely concerns industrial processes, such as the preparation of acetylene and the fluorocarbons, nitrogen fixation, and the reduction of metal ores. In addition, there is some interest in using plasmas to heat inert gases to relatively high temperatures, and to get very high temperatures and very intense heat flows into reactive systems.

To use the energy from a plasma efficiently, the hot reactants must be cooled very rapidly. Further, the production of desirable end-products must involve more than merely trapping the metastable and excited species produced in the hot gas. C. W. Marynowski, R. C. Phillips, J. R. Phillips, and N. K. Hiester [3] have listed two approaches. Firstly, one might control and change the path as a plasma system proceeds from plasma jet temperatures to ambient condition. In this way, normal ambient equilibria are kinetically suppressed in favour of recombination reactions that lead to desirable metastable products. Alternatively, one might react the plasma jet species with an independently introduced reactive medium. This approach could use such techniques as spraying a reactive liquid or solid into the jet, dispersal of the jet gases through a reactive liquid, fluidization of a reactive solid with the jet gases, and transpiration cooling of an expansion nozzle with a reactive liquid.

In the following paragraphs some of the successful plasma applications will be discussed.

Acetylene synthesis. Thermodynamic studies indicate that the stability of hydrocarbons towards their elements at high temperatures increases in the order: alkanes$\ll$alkenes$\ll$alkynes. In addition, species such as CH_3, CH_2, and CH become appreciable constituents of the equilibrium gases. Thus, the decomposition of hydrocarbons at high temperatures would favour formation of certain alkenes and alkynes, either directly or by recombination of such radicals as CH and CH_2. The use of plasma generators for the preparation of acetylene from methane at a reasonable rate requires a temperature of 2000°K or higher, and a quick freezing of the equilibrium mixture to maintain an excess of the rate of formation of C_2H_2 over the rate of decomposition. An analytical study of the decomposition of methane at the high temperatures attainable in a plasma jet reactor was made by J. E. Anderson and L. K. Case [4]. They applied the available thermodynamic and kinetic data so as to predict the overall results of mixing hot hydrogen from a plasma torch with methane. The yield of acetylene is reduced by side reactions which produce diacetylene (C_4H_2) and ethylene (C_2H_4), and these processes compete at high temperatures.

Theoretically, a 90 per cent conversion of methane into C_2H_2 is expected at the quenched products temperature of 1800°K. However, the conversion of acetylene into C_2H_4 and C_4H_2 through side reactions amounts to about 5 per cent for each. An experimental yield of 76 per cent of acetylene was obtained, which is in good agreement with these calculations.

Workers at the Research Institute of Temple University, Philadelphia, have explored C_2H_2 production in a different d.c. plasma jet as well as in an r.f. plasma jet. H. W. Leutner and C. S. Stokes fed CH_4 into an argon plasma and obtained an acetylene yield of 80 per cent based on carbon input. The use of elemental carbon as a consumable anode and hydrogen in a mixture with argon as plasma gas, gave a yield of about 34 per cent. The decomposition of CH_4 in a helium plasma from a radio-frequency plasma torch gave a still lower yield, 29 per cent [5].

Further studies of the C_2H_2 synthesis were made by R. F. Baddour and J. M. Iwasyk [6] who used a high-intensity arc in their investigation of the carbon-hydrogen system. They compared thermodynamic and kinetic data with their experimental results on the basis of the following reactions:

$$C(s) + 2H_2(g) \rightarrow CH_4(g) \quad \ldots .(1)$$

$$2C(s) + H_2(g) \rightarrow 2CH(g) \quad \ldots .(2)$$

$$2CH(g) \rightarrow C_2H_2(g) \quad \ldots .(3)$$

$$H_2(g) \rightarrow 2H \quad \ldots .(4)$$

$$C_2H_2(g) \rightarrow C_2H + H \quad \ldots .(5)$$

While reaction (1) proceeds rapidly at 800°K, reactions (2) and (3) are more favoured at about 2500°K. Reactions (4) and (5) occur in the reverse direction at the quenching temperature. In this case the yield of C_2H_2 is expected to be higher in the quenched gas than in the hot reaction gas. An experimental yield of 24 per cent in the quenched gas, compared with a theoretical acetylene content of 34 per cent, was obtained. This yield is in accordance with the postulate that C_2H radicals are an important intermediate which would give a theoretical yield of 14 per cent in the hot gas.

In all the mentioned syntheses of C_2H_2 in high-intensity arcs or plasma torches, the yield is higher than in any of the previously reported conventional methods, such as pyrolysis of hydrocarbons. The DuPont Company has recently announced the commercial production of acetylene by a 'modified arc process' and the costs appear to be competitive with those for the generation of acetylene by hydrolysis of calcium carbide.

Nitrogen fixation. It is unnecessary to explain the importance of nitrogen fixation as a chemical process. A. V. Grosse and his co-workers at Temple University have used the plasma torch for reactions involving nitrogen, such as the preparation of hydrogen cyanide, cyanogen, nitrogen oxides, and metal nitrides [7, 8].

1. Preparation of hydrogen cyanide (HCN)

This is based on the endothermic reaction

$$2C(s) + H_2(g) + N_2(g) \rightarrow 2HCN(g)$$

Different starting materials have been used and table 1 summarizes the results of the different approaches to the problem. The use of CO as a source of carbon, and NH_3 to provide the required hydrogen and nitrogen, did not lead to successful results.

2. Preparation of cyanogen (C_2N_2)

Cyanogen was prepared in a d.c. plasma torch with graphite as a consumable electrode and nitrogen either as a plasma gas or as a reactant fed into the hot stream of argon plasma. A 15 per cent conversion into C_2N_2 (based on carbon input) was reported. A residence time of the reactants in the plasma reactor of the order of a few tens of milliseconds and a temperature of more than 4000°C were estimated.

3. Preparation of nitrogen oxides

The problem of synthesizing nitric oxide at high temperatures is still a challenge to workers in the field. Thermodynamic and kinetic data predict that temperatures as high as 5000°K still favour nitric oxide formation, but extremely high cooling rates are needed to minimize reformation of the elements. R. C. Philips and F. A. Ferguson [9] calculated the concentration of nitric oxide in the effluent gas from a plasma reactor based on an equimolar flow of nitrogen and oxygen and a reaction temperature of 3500°K. The expected concentration would be 6·6 mole per cent, as compared with 5·1 per cent from processing air. C. S. Stokes and W. W. Knipe reported [7] a conversion of 2 per cent based on oxygen input from a d.c. plasma torch when oxygen in the stoichiometric ratio was fed into a hot stream of nitrogen plasma.

Using a radio-frequency plasma torch, A. V. Grosse, H. W. Leutner, and C. S. Stokes [10] obtained 1–2 per cent yield based on NO_2. The products consisted of 50 per cent NO_2, 40 per cent N_2O_4, and 10 per cent NO. The same workers used modified liquid air at various ratios of nitrogen to oxygen. The $O_2 - N_2$ mixture was fed into a d.c. plasma torch, stabilized by helium. Very small yields of NO and NO_2 (844 and 472 p.p.m.) were obtained, and further work is still needed.

(4) Preparation of metal nitrides

Stokes and Knipe reported [7] the formation of TiN and Mg_3N_2 by feeding powders of the respective metals into a nitrogen plasma. Lustrous crystals of TiN were obtained in a 30 per cent yield. A higher yield, 40 per cent, was obtained in the case of Mg_3N_2. W. L. Jolly has reported, in discussion, the preparation of S_4N_4 by injecting sulphur into a nitrogen plasma and quenching the gaseous products in CCl_4.

Preparation of halocarbons. The preparation of a wide variety of halocarbons is now possible by means of electric arc reactions. J. K. Wolfe, B. Hills, and N. C. Cook obtained CF_4 as the main product of the reaction when a d.c. arc struck between graphite electrodes was immersed in a mixture of charcoal and CaF_2 [11]. CF_3Cl, CF_2Cl_2, $CFCl_3$, and traces of higher fluorochlorocarbons were obtained using an equimolar mixture of $CaCl_2$ and CaF_2. R. F. Baddour and R. B. Bronfin formed C_2F_4 from CF_4 in a high-intensity arc sustained between two graphite electrodes, and A. S. Kanaan and J. L. Margrave synthesized hexachlorobenzene (C_6Cl_6) from CCl_4 vapour exposed to a hot argon plasma [12]. Phillips and Ferguson exposed different halofluorocarbons to an argon plasma jet and quenched the hot stream by bubbling through CCl_4 [3]. A liquid residue of molecular weight 400 was collected but not identified.

Reduction and halogenation of refractory compounds. The high temperatures to which materials can be heated in high-intensity arcs and plasma torches make many refractories reactive. Thermal decompositions, vapour phase halogenations, and carbothermic reductions are among the reactions successfully achieved in this field. S. Korman and his co-workers at Vitro Laboratories have applied the high-intensity arc to several of these processes [13].

Manganese has been successfully extracted from its siliceous ore, rhodonite, by incorporating the $MnSiO_3$ into the carbon anode. Vaporization in the arc decomposes the silicate ore into MnO and SiO_2. The manganese can be recovered from its oxide by selective dissolution of the simple oxides. The method has been suggested for the recovery of such metals as uranium, beryllium, and lithium as well [13].

Carbothermic reductions of metal oxides, such as MgO, have also been attempted. Carbon and MgO were mixed in the anode and the reduction was nearly complete at high temperatures:

$$MgO(s) + C(s) \rightarrow Mg(g) + CO(g)$$

The stability of CO at high temperatures favours complete reduction, but some reformation of the metal oxide occured on cooling. The vapour phase, as confirmed by spectroscopic observations, consisted of elemental Mg, O, and C, and some CO.

When vaporization of an ore-bearing anode takes place in a chlorine atmosphere, the metal oxide (MeO) can be converted quantitatively into the chloride in the vapour phase

$$MeO(s) + C(s) + Cl_2(g) \rightarrow MeCl_2(g) + CO(g)$$

Hydrogen reduction of Al_2O_3 in a plasma jet was investigated, but not thoroughly, by Grosse and his co-workers [10]. Because of the very exothermic nature of the recombination reaction, a high rate of quenching is required. A very poor yield, less than 2 per cent of aluminium, was obtained in experiments conducted in the presence of H_2 or CH_4 as carriers of the alumina powder. Marynowski and his co-workers suggested [3] the possible use of Al_2O_3 for nitrogen fixation

$$5N_2(g) + 2Al_2O_3(s) \rightarrow 6NO(g) + 4AlN(s)$$

Other processes. Other chemical applications of plasma devices involve the interaction of injected vapours, liquids, or solids, with the plasma jet gases. The possibility of completely vaporizing fine powders permits, in turn, the precipitation of an ultra-fine product, potentially of altered stoichiometry. The nickel–nickel oxide system was investigated by T. B. Selover Jr [14]. Ni, either in the form of powder or as volatile $Ni(CO)_4$, was sprayed into the plasma jet confined by a reaction chamber with controlled atmosphere. The nickel fumes produced as a result of the carbonyl decomposition consisted of a physical mixture of nickel and nickel oxide of the formula $NiO_{1.008}$. Analysis showed that the mixture contained 70 per cent NiO and 30 per cent free nickel. The surface area of the collected products was of the order of 150 m^2/g and particle diameters were from 30 to 300Å.

In a radio-frequency discharge, the production of metal oxides by direct oxidation of solid particles or by oxidation of halogen compounds of the metal is now feasible. For example, solid TiO_2 has been prepared from $TiCl_4$ by passing a mixture of $TiCl_4$ and O_2 through a radio-frequency plasma torch [19].

Future applications of plasma torches

Compared with conventional heating devices, both r.f. and d.c. plasma torches present several significant advantages.

1. The temperatures which may be obtained can be considered as practically unlimited.

2. Since the plasma is produced in the form of a jet, its innermost part is extremely hot, while the peripheral parts are at normal temperature. There is, therefore, no contact between the hot gases and a container. This avoids the need for a refractory, and eliminates many problems of contamination.

3. The choice of the plasma gas is almost free—it can be, for example, inert or reactive, a pure gas or a mixture.

4. Reactants can either be passed through the discharge itself or be injected into the slightly cooler regions of an inert or reactive jet of plasma gases.

5. The rate of heat transfer from a plasma to a solid reaches very high values, because of the recombination of atoms, and of ions and electrons, on solid surfaces. This property is very important in heterogeneous reactions.

6. A 10 kW d.c. plasma jet generator with power supply currently costs about $5000. A 10 kW r.f. plasma generator with the induction heater (1–50 Mc/s) currently costs about $10 000. Such prices are not excessive, as no other simple device can generate such high temperatures.

Thus the plasma torch is an ideal tool for conducting homogeneous and heterogeneous flow reactions at very high temperatures. It can be used in two general types of chemical reaction.

Table I

*Summary of data on HCN synthesis in plasma reactors**

Source of constituent elements				Percentage yield based on carbon input		
Carbon	Hydrogen	Nitrogen	Plasma device	HCN	C_2H_2	Remarks
Solid carbon as consumable cathode	H_2 gas fed into the plasma jet	N_2 gas as a plasma gas	d.c. plasma torch	50	13	2% of various hydrocarbons
Solid carbon as consumable cathode	NH_3 gas fed into the plasma jet	N_2 gas as a plasma gas and NH_3	d.c. plasma torch	39	18	
Solid carbon as consumable cathode and CH_4	CH_4 gas fed into the plasma jet	N_2 gas as a plasma gas	d.c. plasma torch	45	6	
CH_4 gas fed into the plasma jet	CH_4 gas	N_2 gas as a plasma gas	d.c. plasma torch with 2% thoriated tungsten electrode	23	62	
CH_4 gas fed into the plasma jet	NH_3 and CH_4 fed in stoichiometric ratio into the plasma jet	N_2 gas as a plasma gas and NH_3	d.c. plasma torch with 2% thoriated tungsten electrode	50	10 to 25	0·8% C_2N_2
CH_4 gas fed into the plasma jet	CH_4	N_2 gas as a plasma gas	r.f. plasma torch	35	24	

* Detailed operating conditions, including gases flow rates and power inputs, are given in the original papers.

The first class includes many endothermic reactions in the temperature range up to 5000–8000°K where molecules are still stable. In these the equilibrium, or pseudo-equilibrium, composition attained at high temperatures can be frozen by quenching the reaction mixture to minimize back-reactions, and several well-known metastable compounds, such as C_2H_2, $(CN)_2$, HCN, and NO_2, have been successfully synthesized with plasma torches. There are, undoubtedly, many endothermic compounds, stable at high temperatures and metastable at room temperature, which are unknown at present because, before the plasma torch, there was no means of synthesizing them. Such preparations represent some of the more exciting goals of plasma chemistry. Endothermic reactions of importance in metallurgy can also be carried out in plasma jets. For example, it is possible to dissociate the elements of a mineral or of a refractory oxide by thermal decomposition, but there is always some recombination on cooling. If this recombination could be prevented, it could lead to great simplification in such metallurgical operations as the extraction of magnesium, aluminium, and sodium from low-grade ores or common rocks and clays and other silicates.

In the second class of reaction, the reaction mixture that is quenched contains species that are completely unstable at normal temperature. In this case the quenching does not freeze an equilibrium composition and the active species recombine. This class is very wide and contains, for example, such reactions as the preparation [9] of organometallics, alkyl silanes, the polymerization of organic radicals [16], and a large group of practically unexplored ion-ion and ion-molecule recombinations. Because of the extremely high energies of such systems when heated above 8000°K, almost any conceivable molecule can be formed by a thermodynamically allowed reaction. Amino acids and other complex species have been synthesized in electric discharges [17] and electron irradiation processes [18] of H—O—C—N systems. Many other products would be produced if plasma species could be so manipulated as to follow selected kinetic paths.

The complications of plasma-jet reactions are such that they cannot be understood unless a number of very rough assumptions are made and this difficulty is, of course, a hindrance in finding a successful application. Let us consider the decomposition of CH_4 in a plasma reactor. This has been treated independently in two recent papers with completely different interpretations [4, 19]. In the first, it is assumed that the rate of mixing of the plasma with CH_4 is infinite; in the second, the rate of reaction is considered as diffusion-dependent, If such fundamental questions are difficult to resolve, then more complicated problems, such as the control of ion recombination reactions to yield favourable products, are still a challenge to the chemist.

In several promising fields, such as metallurgy, the plasma chemist is faced with an additional problem. The short contact times for reactants in plasma jets are not favourable for heterogeneous reactions. For example, in the hydrogen reduction of a metal oxide, the molten metal often forms a protective layer which prevents complete reaction in the hot gases. Further, in the reduction of metal halides by hydrogen, the metal produced is often poorly crystallized and sometimes is pyrophoric. The major metallurgical applications, therefore, utilize the plasma supply mainly to heat solids to 2000–4000°K.

A final problem of application is that industrial organizations tend to be cautious in their support of a field where the results are hardly predictable. There is a temptation to try first to see if existing processes can be replaced advantageously by plasma-jet processes, and thus to ignore more imaginative ideas.

References

[1] Giannini, G. M. *Sci. Amer.*, **197**, 80, 1957.

[2] Reed, T. B. *Int. Sci. Technol.*, No. 6, 42, June 1962.

[3] Marynowski, C. W., Phillips, R. C., Phillips, J. R., and Hiester, N. K. *Ind. Engng. Chem.*, (*Fundamentals*), 1, 52, 1962.

[4] Anderson, J. E. and Case, L. K. *Ind. Engng. Chem.*, (*Process Des. Devel.*),1 , 161, 1962.

[5] Grosse, A. V., Stokes, C. S., Cahill, J. A., and Correa, J. J. 'Plasma Jet Chemistry' (Final Report). Office of Naval Research (NR 052-429). 30 June 1963.

[6] Baddour, R. F. and Iwasyk, J. M. *Ind. Engng Chem.* (*Process Des. Devel.*), 1, 169, 1962.

[7] Stokes, C. S. and Knipe, W. W. *Ind. Engng Chem.*, **52**, 287, 1960.

[8] Leutner, H. W. *Ibid.*, (*Process Des. Devel.*), 1, 166, 1962.

[9] Phillips, R. C. and Ferguson, F. A. 'Proceedings of an International Symposium on High Temperature Technology'. P. 192. McGraw-Hill, New York. 1959.

[10] Grosse, A. V., Leutner, H. W., and Stokes, C. S. 'Plasma Jet Chemistry' (First Annual Report). Office of Naval Research (NR 052-429), 31 December 1961.

[11] Wolfe, J. K. and Cook, N. C. U.S. Patent 2,835,711. 20 May 1958.

[12a] Kanaan, A. S. and Margrave, J. L. *Int. Sci. Technol.*, No. 875, August 1962.

[12b] Baddour, R.F. and Brontin, B.R. *Chem. Engng. News*, 35, April 22, 1963.

[13] Harris, V., Holmgren, J., Korman, S., and Sheer, C. *J. electrochem. Soc.*, 106, 874, 1959.

[14] Selover, T. B. Jr. Preprint 71. 'Symposium on Flames and Plasmas in Chemical Processing'. American Institute of Chemical Engineers Meeting, New York, December 1961.

[15] 'TiO_2 by Radiofrequency Plasma Torch', B.P. 926,716 to DuPont, as cited by Monroe, A. G. *Proc. Int. Cong. High Temp. Technol.*, 76, 1963.

[16] Farlow, M. W. U.S. Patent 2,709,192. 24 May 1955 (to E. I. DuPont).

[17] Miller, S. L. *J. Amer. chem. Soc.*, **77**, 2351, 1955.

[18] Ponnamperuma, C., Lemmon, R. M., Mariner, R., and Calvin, M. *Proc. nat. Acad. Sci., Wash.*, **49**, 737, 1963.

[19] Freeman, M. P. and Skrivan J. F. 'Rates Studies of the Decomposition of Ammonia and Methane in a Plasma Jet'. American Institute of Chemical Engineers Meeting, New York. December 1961.

Electron microscopy of the cell surface

E. G. Gray

Modern staining and sectioning have increased the resolution of the electron microscope to 20Å, and these techniques reveal that the surface of the cell is multilayered. These layers may be significant in synaptic contacts, in the inhibition that probably prevents the cancerous growth of cells, and in the protection of the brain against poisonous substances by the blood-brain barrier. This article describes the structure of protoplasmic membranes and discusses its significance.

The electron microscope has given useful information about cell structure for about thirty years, but it is only during the last decade that techniques of fixation, heavy-metal staining, and ultra-thin sectioning of material embedded in plastic have made possible a really penetrating scrutiny of cell organelles at magnifications up to and exceeding one hundred thousand times linear. Structural detail in particles, fibrils, and membranous interfaces can now be examined with a resolution of 20 Å, compared with the 2000 Å of the light microscope. The cell surface, for long inaccessible to the morphologist, is now being carefully studied, and the special layering of the proteins, polysaccharides, and lipids that for long have been thought to form the basic architecture of cell membranes is now being successfully investigated. Cells may lie so as to be carried in extracellular material such as blood, lymph, or tissue fluids, for example, but they are commonly found in aggregates, often massive, forming the tissues of the body. The electron microscope reveals how the cell surfaces become more or less intimately related, either temporarily or permanently. In recent years, cancer workers have shown increasing interest in cell contacts, for the malignant transformation seems to be directly related to the cell's changed contact properties.

The cell surfaces are also of great interest in the nervous system, for there they have special regions of apposition, the synapses. The electron microscope shows that these are very complex structures and an understanding of their morphology may eventually provide further clues to the mechanisms underlying the higher functions of the nervous system. This article will be confined to comments on the electron microscopist's approach to all of these topics, although this is not the only way, of course, of investigating cell-surface morphology.

The concept of the unit membrane

Physico-chemical evidence suggested [1, 2] that certain cell-surface membranes consisted of bimolecular leaflets of lipid with the polar groups orientated outwards. Each polar surface was faced by a protein monolayer (figure 1). Early electron microscopy, using osmium tetroxide, then the most common fixative, revealed a dense lamina in sections of the cell surface, but this usually appeared as a single dense line 60–80 Å thick (figure 3b) and so gave no clue to the supposed underlying molecular lamination. Using the same fixative, H. Zetterquist [3] observed in 1956 a trilaminated structure in restricted regions of surface membranes of intestinal cells. This was seen in sections as two dense lines with a clear zone between, although other regions showed the usual single dense line. In the same year, J. H. Luft [4] described potassium permanganate as a new fixative for electron microscopy.

J. D. Robertson [5, 6] was the first to use permanganate as a fixative in a high-resolution study of a variety of cell membranes with the electron microscope. This method consistently delineated sections of the surface membrane not as a single dense line, but as a triple structure, consisting of two dense lines, each ~25 Å thick enclosing a pale zone, also ~25 Å wide, giving a total thickness of some 75 Å. Figure 3(*a*) shows portions of three adjacent cells. The central one is narrow and separates the other two. The triple structure can be seen bordering each of their surfaces.

Robertson suggested that permanganate fixation gives a true reflection of the underlying molecular architecture of the membrane and termed the triple structure the 'unit membrane'. The permanganate reacts in such a way that it becomes bound, probably along the regions of protein attachment to the polar groups (figure 1). The total width of the unit membrane, 75 Å, would give just sufficient room in the pale zone between the two laminae to enclose the bimolecular leaflet of lipid. The pale zone appears structureless in the electron beam because the non-polar regions of the lipid molecules that it contains are unreactive with the permanganate fixative.

Before electron microscopy, much of the evidence for the ultrastructure of layered lipo-protein systems of cells was obtained indirectly by polarization microscopy and X-ray diffraction studies. The myelin sheath of nerve fibres and the outer segments of the rods and cones are well known examples of lamellated structures. The myelin sheath, in particular, has now become of special relevance to the study of cell membranes, for B. B. Geren [7] has shown that as nerve fibres grow in the

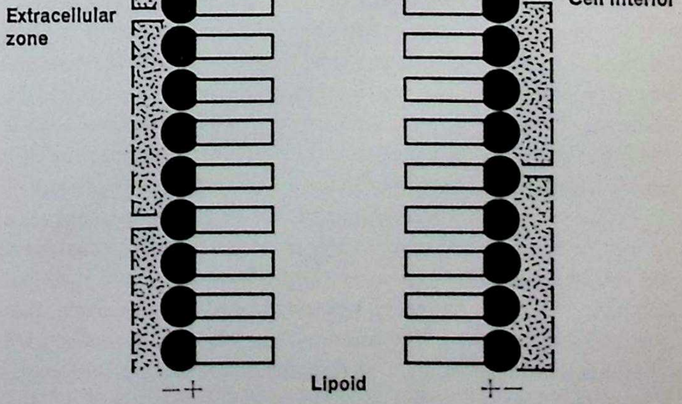


Figure 1 General architecture of the cell membrane.

E. G. Gray, *Ph.D.*
Was born in Abergavenny in 1924. He was educated at University College of Wales, Aberystwyth and is at present Reader in Cytology at University College, London. His special interest is electron microscopy of the nervous system.

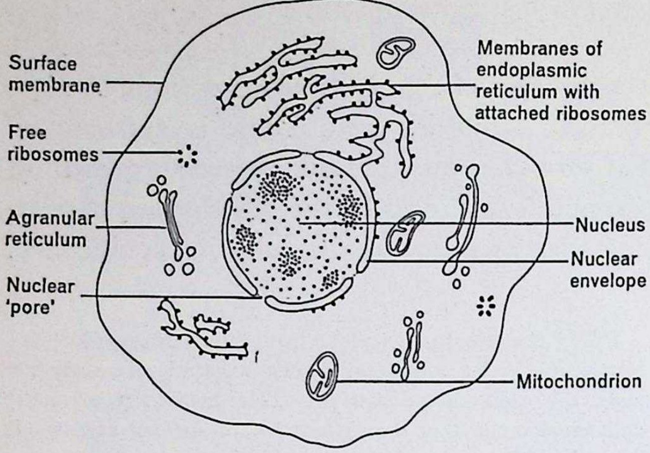


Figure 2 The generalized animal cell from electron microscopy.

embryo, the Schwann cells wrap themselves round and round the fibres, squeezing out the protoplasm and leaving a compact lamellated sheath. Robertson [6] made a detailed study of this sequence of events with the electron microscope, using permanganate as a fixative, and showed clearly how the unit membrane of the Schwann cell can be followed into, and seen to be the sole constituent of, the helical lamellae of the myelin (figure 4). So the myelin is, in fact, a series of apposed cell membranes all lying parallel and hence capable of analysis by polarization optics or X-ray diffraction. These results concerning the molecular dimensions and spacing can be compared with the direct observations of unit membranes by electron microscopy. There is, in fact, a good agreement in these calculations, allowing for the shrinkage involved in fixing, dehydrating, and other processing for electron microscopy.

Membranes inside the cell

In addition to the surface membrane, a quite complicated system of internal membranes is revealed in the cell by the electron microscope (figure 2). There is a system of interconnected sacs, the endoplasmic reticulum, which expands to form an envelope of paired membranes around the nucleus. Small dense particles, termed ribosomes because they contain ribose nucleic acid, adhere to these membranes and play an important part in protein synthesis. A second set of membranes, termed the agranular reticulum because the membranes have no adherent ribosomes, coincides, at least in position, with the Golgi apparatus that can be seen by light microscopy after special staining. In addition, mitochondria are scattered through the cytoplasm, and these have two membranous sacs, one inside the other, the internal one having inwardly projecting crests. The mitochondria have various enzymes that lead to the production of energy-rich adenosine triphosphate, which then freely diffuses to meet energy requirements in various parts of the cells. Robertson [5, 6] showed that all these membrane systems show the triple-membrane structure under the electron microscope, after fixation with permanganate, and it may therefore be concluded that they all have a basic molecular architecture comparable with that of the surface membrane.

What then is the present-day status of the unit-membrane concept? There is still controversy over the correlation between the single dense line seen in sections of osmium-fixed material and the triple structure of permanganate preparations. Not infrequently, the triple structure can also be demonstrated with osmium fixation. This variability is not really understood. Also, careful measurements have revealed that, in fact, the 'unit membranes' of the endoplasmic reticulum and mitochondria are only 85–90 per cent of the thickness of the surface membrane [8]. This as yet presents no serious challenge to the theory, for it could well reflect a true structural difference. Many workers have now demonstrated, with permanganate, the triple structure in a wide variety of cells from almost all groups of animals and some plants and bacteria. Even certain viruses, having successfully parasitized the interior of a cell, encase themselves within a small portion of the unit membrane as they move out through the cell surface [9].

Open and closed cell contacts

Permanganate fixation, then, has two especially important advantages. Firstly, it can show whether a particular membrane has a lamellated protein-bimolecular lipid structure or not. Secondly, it can locate the outer surface of the membrane exactly, so that when the surface membranes of two cells are apposed at a certain point, the exact extent of the intervening channel can be measured, or, alternatively, a surface-to-surface contact may be revealed to be so intimate that there is no resolvable gap at all. This sort of information is important in studying cell-surface relations, but we also need to know about other structures at the cell surface, which might indicate regions of special function or points of strong mechanical attachment. So nowadays, as can be seen from the illustrations here, it is becoming the practice to study membranes by the two separate techniques: permanganate fixation for the 'molecular skeleton', and osmium tetroxide fixation for surface specializations. For either method, contrast in the electron beam is further enhanced by 'staining' the fixed tissue with heavy metals, actually compounds of uranium, tungsten, or lead.

Let us now turn to further examples of specialization of the surface membrane (figure 3a). There is an obvious extracellular zone between the upper pair of apposed unit membranes (permanganate-fixed) shown in section. It appears structureless and varies from 100–200 Å across. We would very much like to know the exact width of this cleft in the living cell before fixation, but such information is not as yet accessible by direct observation. The lower pair of unit membranes show a cleft on the left, but at region x, the membranes come together, and so the extracellular channel is apparently completely closed, to form a complex about 150 Å thick. This structure is described by the terms 'closed contact' or close contact, external compound membrane, or quinti-laminar structure—the latter because there are three dense lines separated by two pale zones. For comparison, a section of a closed contact between two smooth muscle cells is shown (figure 3c, [10]) in an osmium-fixed preparation. The widely spaced membranes on the left show no triple—unit membrane—structure, but at x they come together and form a closed contact directly comparable with the permanganate picture of the closed contact (figure 3a). The outer surfaces of the membrane, therefore, seem to be well preserved in osmium-fixed preparations only when they are in intimate contact and hence more stable in the fixative [6].

In the last few years, closed contacts have been discovered in several tissues. In addition to the examples just mentioned, there are the membrane contacts formed from the Schwann cell surface in the formation of the myelin sheath (figure 4): this serves as an insulator, and so greatly increases the efficiency of conduction in the nerve fibre it surrounds. Closed contacts are invariably present between adjacent endothelial cells that line blood vessels and between other sorts of epithelia. They probably play a role in adhesion (see below) but may also prevent the leakage between the cells of fluid containing ions and molecules.

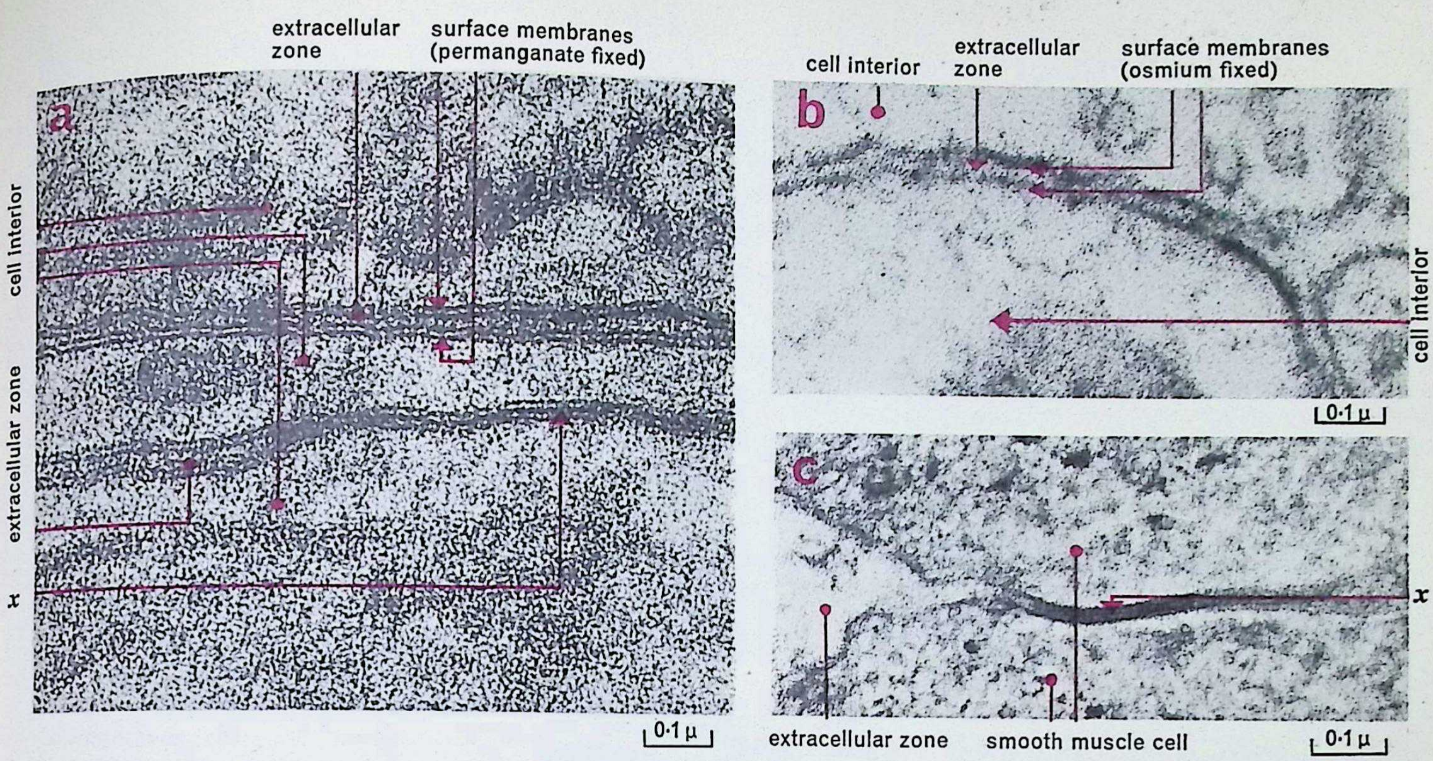


Figure 3 (a) Three cellular laminae with (above) membranes showing open contact, and (below) closed contact ($KMnO_4$ fixation) (b) Apposed surface membranes with open contact (OsO_4 fixation) (c) Closed contacts between two smooth muscle cells (OsO_4 fixation [10]).

In the brain there is a layer of special cells, the neuroglia, that form sheaths around the blood vessels and isolate them from the brain. The brain has the special property of being protected by a mechanism which selectively prevents a variety of toxic agents, but unfortunately not all, from entering it from the blood stream. This is known as the blood-brain barrier. Probably the neuroglia play an important role in deciding what passes to the brain. The fine channels running between these neuroglia are sealed by closed contacts [11], so probably nothing can pass into the brain without actually passing through the neuroglial cytoplasm for censorship. Finally, closed contacts may play another quite different role, for there is evidence that electrical excitations may pass from one cell to another at these points. This is considered in detail below.

Desmosomes or attachment plaques form another type of membrane specialization, especially between cells subjected to continuous mechanical distortion, as in the skin (figure 5). Four desmosomes can be seen along the apposed convoluted surfaces; the unspecialized regions are marked by arrows. At the desmosome, the cleft widens and protein fibrils form a tangled dense mass (f) where they are attached to the cytoplasmic surfaces of the membranes. The fibrils (termed tonofibrils) trail out on each side, anchoring the membrane plaques firmly to the cytoplasm. The cleft of the desmosome contains a sheet of dense material (cm), which probably has a cementing function. In sections of permanganate-fixed desmosomes, the dense tonofibrillar masses and cement substance are invisible in the electron beam, and the unit membranes can be clearly seen lining the cleft and corresponding in position to the outer border of the dense tonofibrillar mass.

So far then, we have seen three types of contact. First, there is the more usual open contact where the cell membranes lie apposed, with a 'gap' of about 150 Å between, and no other specializations. These regions probably have no strong adhesion properties. Secondly, there is the closed contact, where the membrane surfaces are touching and there may even be interweaving of the outer molecular layers. There is some

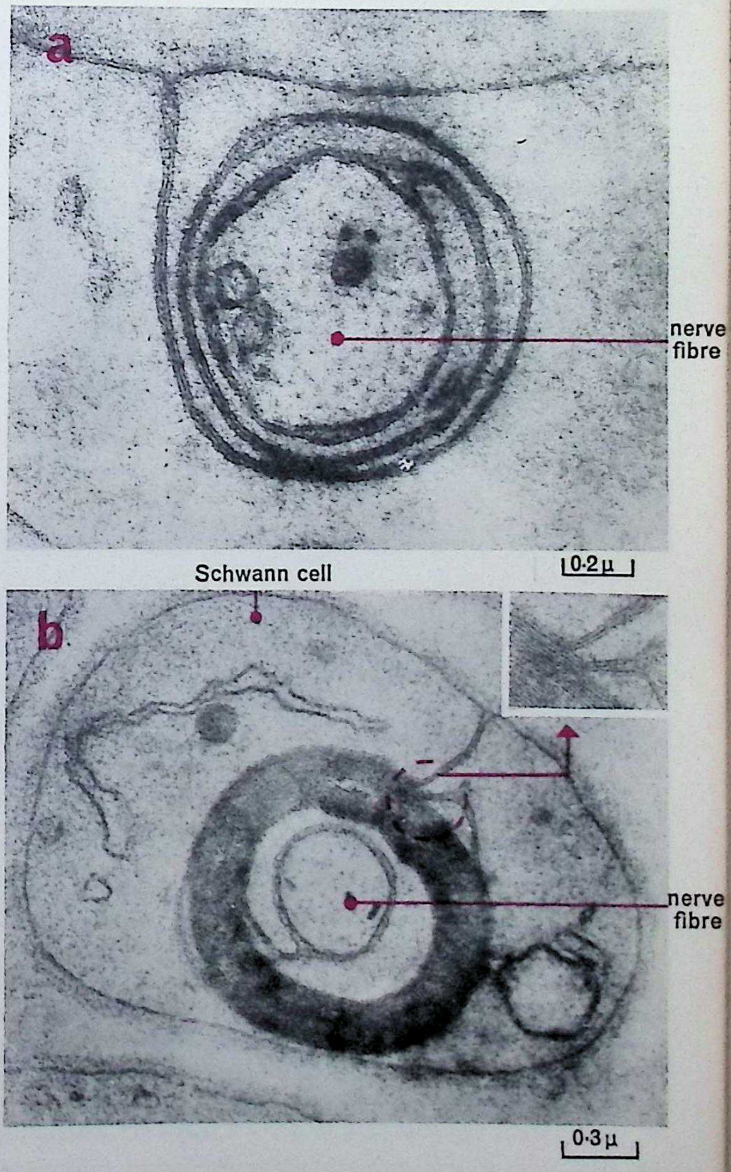


Figure 4 Myelin formation from the Schwann cell surface membrane [6]. (a) The Schwann cell is folding round the nerve fibre in the embryo, (b) Compact myelin sheath formed in the adult. (By kind permission of Dr J. D. Robertson).

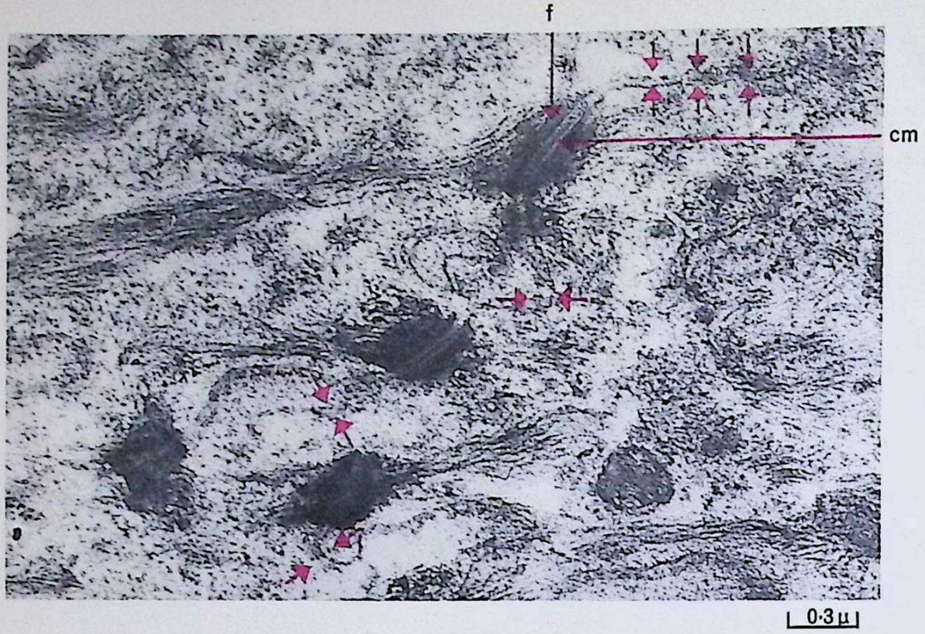


Figure 5 Desmosomes in guinea-pig skin. cm: dense material that probably has a cementing function. f: tangled mass of fibrils. The pairs of arrows indicate the path of the boundary between the cells.

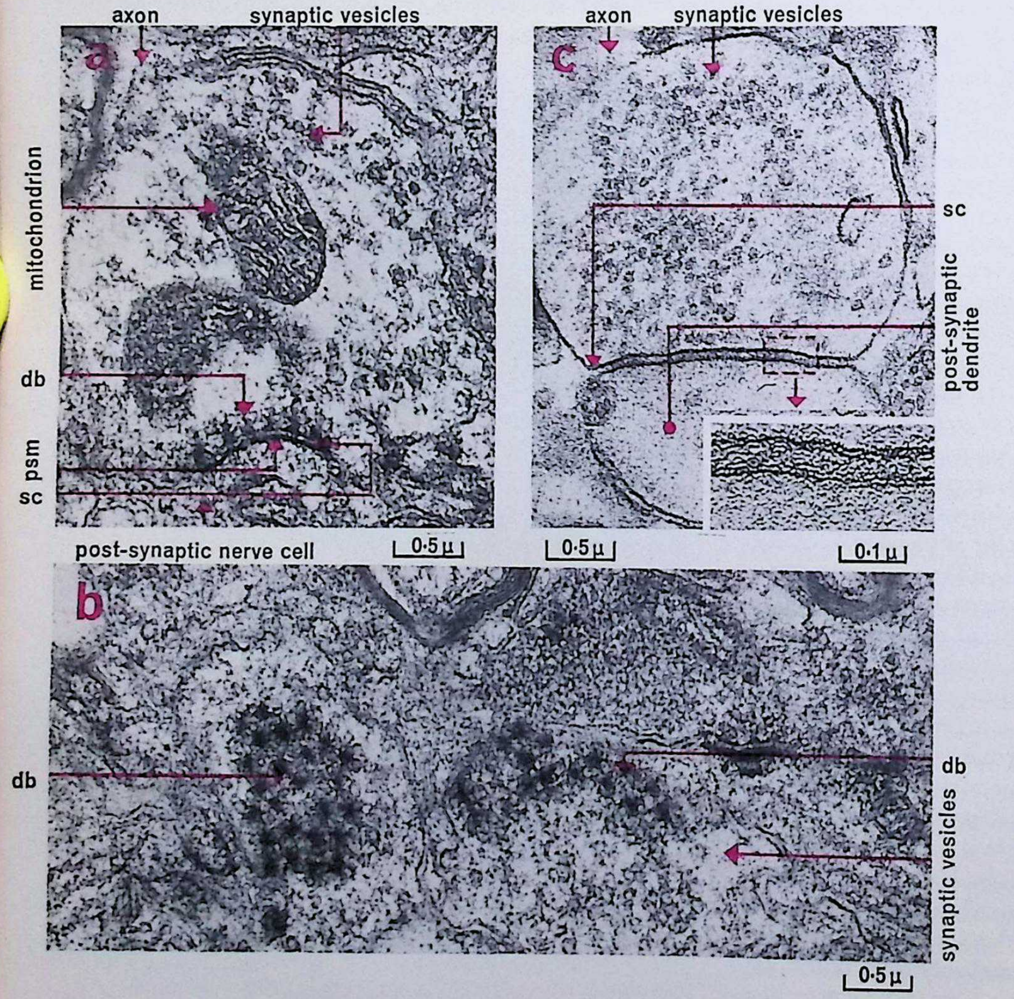

axon
synaptic vesicles
a
mitochondrion
db
psm
sc
post-synaptic nerve cell
0·5 μ
axon
synaptic vesicles
c
sc
post-synaptic dendrite
0·5 μ
0·1 μ
b
db
db
synaptic vesicles
0·5 μ

Figure 6 Synapses: (a) and (b) OsO_4-fixed [19], (c) $KMnO_4$-fixed.

evidence that these are points with special adhesive properties [12], but more work is required on this point. In the third type, the desmosomes are definitely points of especially strong adhesion. Here the membranes are anchored by tonofibrils, and the extracellular cleft is rather wide and contains a cementing substance.

Cancer and cell contacts

The malignant transformation is primarily seen as a rapid and unceasing multiplication of a cell population, followed by an invasive phase. Non-malignant cells, after they divide, move into position to form the adult tissue, and when they touch each other, their movement becomes inhibited. This is the phenomenon of contact inhibition that was described by M. Abercrombie and J. E. M. Heaysman [13]. It is probably of great significance that cancer cells, on the contrary, show no such movement inhibition when they touch [14]. The electron microscope may be able to detect differences between malignant and non-malignant cells in the ways that they meet and contact each other.

There is evidence that electrical impulses may pass across intimately apposed regions of cell surfaces, and it could be that this sort of interaction between the two cells may cause movement inhibition. Such evidence stems from a study of certain rather uncommon synapses (for example, in the crayfish and on the large Mauthner cells of goldfish [15]) and of special contacts between smooth muscle cells (figure 3c). Direct electrical transmission, as opposed to the release of chemical transmitter, is known to take place at these points and electron microscopy reveals that they are, in fact, closed contacts [16, 10]. R. Devis and D. W. James [17] have observed that when mobile, non-malignant fibroblast cells touch each other—resulting in movement inhibition—they may form closed contacts.

However, although mature inactive cells in normal tissues may not show closed contacts between them, they have all presumably come to rest as a result of contact inhibition. It remains to be seen whether cells in general make 'closed contacts' when first they meet, with the membranes later disengaging and moving apart slightly to form the more usual 150 Å cleft.

Synapses

One day, perhaps not so very far ahead, the underlying mechanisms in the brain responsible for the highest human functions—imagery, learning, recollection, insight, and so on—will be understood. Already, with the aid of the electron microscope, we can see the minutest details of brain structure, and the final analysis may well depend on understanding the morphology of the special cell contacts, the synapses.

Nerve fibres connect, in a highly complex way, the various nerve cells of the brain to one another. They conduct messages, coded in the form of groups of electrical impulses. Once the impulse starts at one end of the fibre it proceeds to the other end with certainty. The synaptic junctions, in contrast, are regions of uncertainty. The message may or may not get across to the succeeding nerve cell. Usually the message will be modified and in a complex way related to all the present and past activity of other synapses feeding into the nerve cell. The core of the problem is to understand this interaction both functionally and morphologically. In the human brain there are approximately 10^{10} nerve cells and one cell may have as many as ten thousand synaptic contacts on it [18].

At present, we are trying to use the electron microscope to understand the morphology of the synapse. There is much physiological evidence to suggest that, at most synapses, the nerve impulse when it comes along, liberates small packets of a chemical 'transmitter'—acetylcholine at some synapses—into the synaptic cleft. The chemical diffuses in solution across the cleft in a fraction of a millisecond, and initiates electrical changes in a region of the postsynaptic nerve cell. Thus the cell contacts that are specialized as synapses must be so constituted that (*a*) the contact is made not only with the correct cell but with a specific region of that cell, and (*b*) packets of chemical transmitter can be liberated from the presynaptic membrane into the synaptic cleft when an impulse comes down the nerve fibre (the axon).

Electron microscopy reveals numerous so-called synaptic vesicles (figures 6a and 6c) in the expanded tip of the axon, the presynaptic bag. There is some evidence to suggest that these contain the transmitter substance and that they move towards the synaptic cleft (sc) and discharge their contents there when an impulse arrives. Unfortunately, the electron microscope helps little in revealing this mechanism. Figure 6c shows a section of the synapse fixed with permanganate. The unit membranes of the cleft are shown in the inset. For comparison, a similar synapse is shown in figure 6a fixed with osmium and with the contrast increased with phosphotungstic acid. This second method shows that a series of regularly spaced dense bodies (db) project out from the presynaptic membrane [19], and when cut in the other plane (figure 6b), these can be seen to have a regular spacing. The postsynaptic membrane (psm) shows a continuous dense deposition along it. Thus, using the appropriate fixing and staining technique, the synaptic membrane contact region shows great complexity. These dense projections, at very high magnifications, appear to consist of small coils of filaments (seen here as sections cut in planes depending on their orientation), which probably perforate the presynaptic membrane at these regularly arranged dense points. The perforations may then constitute the channels for the liberation of the transmitter chemical when the stimulus arrives. The synaptic vesicles are sometimes, but not invariably, aggregated near these dense projections and it remains to be seen whether, in fact, they do contain and carry the transmitter towards the presynaptic membrane. In short, it is no poetic exaggeration to say that somewhere here lurk important clues to the nature of the mind of man.

References

[1] Harvey, E. N. and Danielli, J. F. *Biol. Rev.* **13**, 319, 1938.

[2] Davson, H. and Danielli, J. F. 'The permeability of natural membranes'. Cambridge University Press, London. 1952.

[3] Zetterquist, H. 'The ultrastructural organization of the columnar absorbing cells of the mouse jejunum'. Thesis. Karolinska Institutet, Stockholm. 1956.

[4] Luft, J. H. *J. biophys. biochem. Cytol.*, **2**, 799, 1956.

[5] Robertson, J. D. *Biochem. Soc. Symp.*, **16**, 3, 1959.

[6] *Idem*, in 'Molecular Biology', p. 87. Academic Press, New York. 1960.

[7] Geren, B. B. *Exp. Cell Res.*, **7**, 588, 1954.

[8] Yamamoto, T. *J. Cell Biol.*, **17**, 413, 1963.

[9] Epstein, M. A. *Brit. med. Bull.*, **18**, 183, 1962.

[10] Evans, D. H. L. and Evans, E. I. *J. Anat., Lond.* **98**, 37, 1964.

[11] Gray, E. G. in 'Electron microscopy in Anatomy'. Edward Arnold, London. 1961.

[12] Whittaker, V. P. and Gray, E. G. *Brit. med. Bull.*, **18**, 223, 1962.

[13] Abercrombie, M. and Heaysman, J. E. M. *Exp. Cell Res.*, **6**, 111, 1953.

[14] Abercrombie, M. and Ambrose, E. J. *Cancer Res.*, **22**, 525, 1962.

[15] Furshpan, E. J. and Potter, D. D. *J. Physiol.*, **145**, 289, 1959.

[16] Robertson, J. D., Bodenheimer, T. S., and Stage, D. E. *J. Cell Biol.*, **19**, 159, 1963.

[17] Devis, R. and James, D. W. *Nature, Lond.*, **194**, 695, 1962.

[18] Armstrong, J. and Young, J. Z. *J. Physiol.*, **137**, 10P, 1957.

[19] Gray, E. G. *J. Anat., Lond.*, **97**, 101, 1963.

Permafrost and related engineering problems

R. J. E. Brown and G. H. Johnston

Roughly one-fifth of the world's land area is subject to permafrost, that is, perennially frozen ground. It is particularly important in the Soviet Union and in Canada, each of which has roughly half its territory affected. Permafrost not only poses questions of great geophysical interest but presents the civil engineer with many exceedingly difficult problems. This article reviews the principles governing the distribution of permafrost and some of the ways of overcoming the difficulties.

The term 'permafrost' was coined in 1943 by S. W. Muller, who used it as a convenient short form of 'permanently frozen ground' [1]. It describes the thermal condition of earth materials such as rock and soil when their temperature remains below 0°C continuously for a number of years, which may be as few as two or as many as tens of thousands. In recent years, however, it has been realized that permafrost is not necessarily permanent. Changes in climate and terrain can cause the permafrost to thaw and disappear, and the term 'perennially frozen ground' is now used instead of 'permanently frozen ground'.

Permafrost underlies 20 per cent of the world's land area, being widespread in North America, Eurasia, and Antarctica. In the northern hemisphere, it occurs mostly in Canada and the Soviet Union, each country having about one-half of its total land area underlain by it (figure 3) [2, 3].

The origin of permafrost is not well understood, but it is suspected that it first appeared during the cold period at the beginning of the Pleistocene. During subsequent periods of climatic fluctuations, corresponding changes have occurred in its extent and thickness. It is known at present to be diminishing in some areas and increasing in others.

Distribution of permafrost and thermal considerations

In the northern part of the area shown in figure 3, permafrost is continuous, but further south it is discontinuous. In the continuous zone, it occurs everywhere under the ground surface and may be hundreds of metres thick; in the discontinuous zone, permafrost exists in combination with areas of unfrozen material (figure 1). In the southern area of the discontinuous zone, it occurs as scattered patches and is only a few metres thick. The thickness of permafrost at any particular point is determined by the close and complex interaction of a large number of climatic and terrain factors. The most important of these factors are air temperature, relief (slope and aspect), vegetation, drainage, snow cover, and soil type.

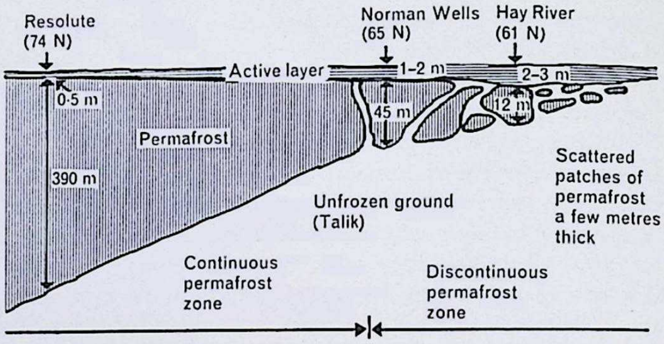

Resolute
(74 N)
Norman Wells
(65 N)
Hay River
(61 N)
Active layer
1-2 m
2-3 m
0·5 m
45 m
12 m
Permafrost
390 m
Unfrozen ground
(Talik)
Scattered
patches of
permafrost
a few metres
thick
Continuous
permafrost
zone
Discontinuous
permafrost
zone

Figure 1 Typical vertical distribution and thickness of permafrost.

Above the permafrost is a surface layer of soil or rock, called the 'active layer', which thaws in summer and freezes in winter (figure 2). Its thickness depends on the same climatic and terrain features as affect the permafrost. The thickness may vary regionally from several metres in the southern fringe of the discontinuous zone to only a few centimetres in the north of the continuous zone. Local variations in terrain factors may cause the active layer to vary in thickness from less than one metre to several metres within any one area.

In permafrost regions, the mean temperature decreases steadily from the ground surface to a depth ranging approximately from 10 to 30 metres. Below this depth, the temperature increases steadily under the influence of the heat from the Earth's interior. After a time-lag determined by depth and local conditions, fluctuations in air temperature during the year produce corresponding temperature fluctuations about the mean annual ground temperature to depths of some 20 metres. The amplitude of these fluctuations decreases with depth to less than 0·1°C. The depth at which fluctuation becomes imperceptible is called the 'level of zero annual amplitude'. Below this, ground temperatures change only in response to long-term climatic changes extending over centuries.

There is a broad relationship between mean annual air temperature and mean annual ground temperature [4]. Observations at various places have shown that the mean

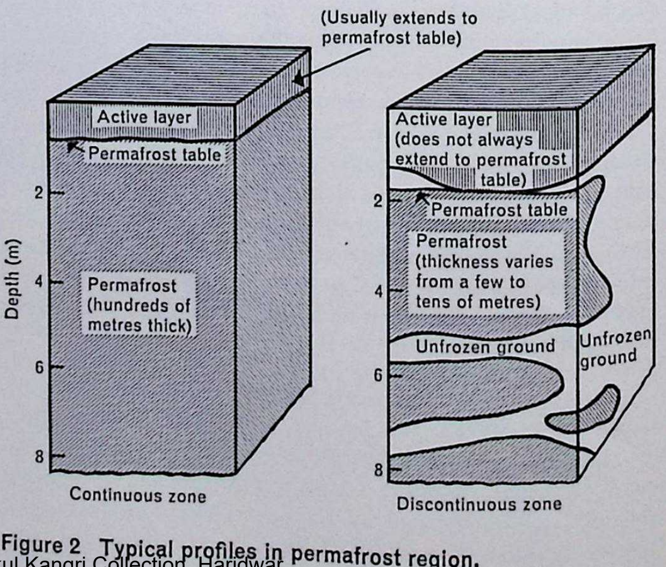

(Usually extends to
permafrost table)
Active layer
Permafrost table
Permafrost
(hundreds of
metres thick)
Depth (m)
2
4
6
8
Continuous zone
Active layer
(does not always
extend to permafrost
table)
Permafrost table
Permafrost
(thickness varies
from a few to
tens of metres)
Unfrozen ground
Unfrozen
ground
Discontinuous zone

Figure 2 Typical profiles in permafrost region.

R. J. E. Brown, *M.A., Ph.D.*
Was born in Toronto in 1931 and educated at the University of Toronto and Clark University, Massachusetts. In 1953 he joined the National Research Council, Canada, and has since been attached to the Northern Research Group, Soil Mechanics Section, Division of Building Research. A geographer, he has been concerned with mapping the distribution of permafrost in Canada and investigating the physical factors affecting this.

G. H. Johnston, *B.Sc.*
Was born in 1927 in Winnipeg and is a civil engineering graduate of the University of Manitoba. He is attached to the same group of the National Research Council as Dr Brown, and has been particularly concerned with site investigation and construction techniques in permafrost areas and the effect of permafrost on engineering structures.

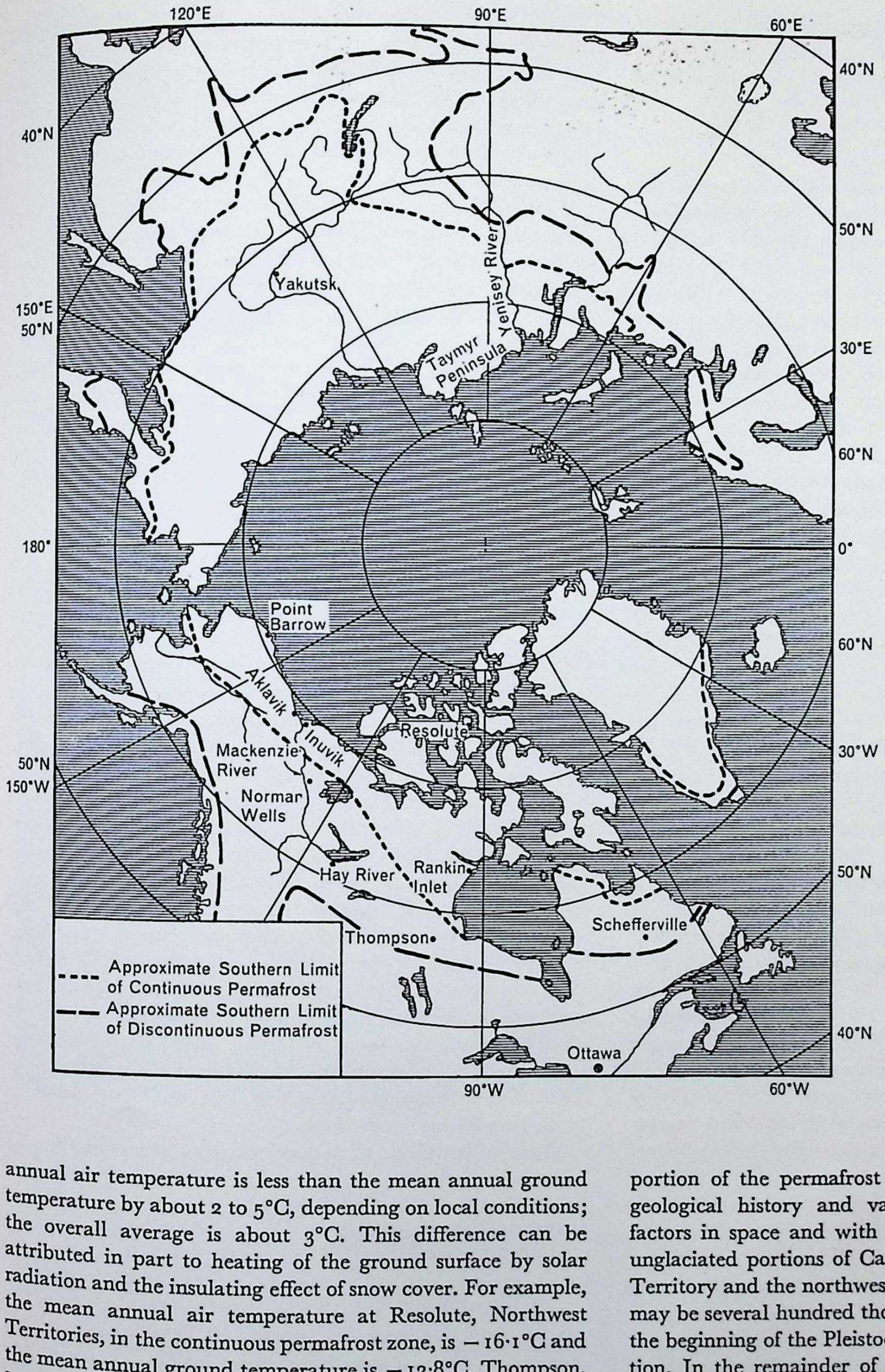

120°E
90°E
60°E
40°N
40°N
150°E
50°N
50°N
30°E
60°N
180°
0°
60°N
30°W
50°N
150°W
50°N
40°N
90°W
60°W
Yakutsk
Yenisey River
Taymyr Peninsula
Point Barrow
Aklavik
Inuvik
Resolute
Mackenzie River
Norman Wells
Hay River
Rankin Inlet
Thompson
Schefferville
Ottawa
Approximate Southern Limit of Continuous Permafrost
Approximate Southern Limit of Discontinuous Permafrost

Figure 3 Map showing distribution of permafrost in the northern hemisphere, and position of places mentioned in text.

annual air temperature is less than the mean annual ground temperature by about 2 to 5°C, depending on local conditions; the overall average is about 3°C. This difference can be attributed in part to heating of the ground surface by solar radiation and the insulating effect of snow cover. For example, the mean annual air temperature at Resolute, Northwest Territories, in the continuous permafrost zone, is −16·1°C and the mean annual ground temperature is −12·8°C. Thompson, Manitoba, in the discontinuous zone, has a mean annual air temperature of about −4°C and a mean annual ground temperature in the permafrost of about −0·5°C.

Over a long period of time, a change in the mean annual air temperature can result in a significant change in the extent and thickness of permafrost. Geothermal gradients ranging from about 1°C/20m to 1°C/160m—depending to some degree on the type of soil or rock—have been observed in permafrost regions in Canada, Alaska, and the Soviet Union, A change of 1°C in the mean annual air temperature, for example, could result, over a long period of time, in a change of 1°C in the mean annual ground temperature. This would cause a change in permafrost thickness of approximately 20 to 160 metres.

The age of permafrost may differ considerably from one portion of the permafrost region to another, depending on geological history and variations in climatic and terrain factors in space and with time. For example, permafrost in unglaciated portions of Canada, such as the Western Yukon Territory and the northwest portion of the arctic archipelago, may be several hundred thousand years old, having formed at the beginning of the Pleistocene, before the continental glaciation. In the remainder of the country, covered by ice, it is unlikely that existing permafrost formed until after the final retreat of the ice sheet, ten to twenty thousand years ago. In areas inundated by postglacial lakes and marine transgression, however, permafrost formation would not have begun until these bodies of water receded, several thousand years after the ice retreated.

An estimate of the time required to form thick masses of permafrost can be obtained by simple calculations if some assumptions are made. For example, it can be calculated, on the basis of simple conduction theory, that at Resolute about 10 000 years was required for the ground to freeze to a depth of 390 metres, the present estimated thickness of the permafrost. A more precise estimate would have to take account of such factors as the latent heat of fusion of ice, proximity of the ocean, fluctuations in mean annual air temperature since the initiation of permafrost accumulation, and changes in terrain conditions. Nevertheless, the above value must be of the right order of magnitude.

Terrain influences

Although climate controls the broad pattern of permafrost distribution, other factors are necessary to explain local variations. Climatic differences alone do not explain the variations in thickness and temperature at depth in neighbouring areas of the continuous zone that have similar mean annual air temperatures. Nor do they explain the patchy occurrence of permafrost in the southern fringe of the permafrost region. These variations in permafrost occurrence appear to be governed predominantly by local variations in micro-climate and such features of the terrain as relief, drainage, vegetation, snow cover, and soil type.

Relief influences the amount of solar radiation received by the ground surface. In the discontinuous zone, this may result in permafrost occurring on north-facing slopes but not on adjacent slopes facing south. In the continuous zone, permafrost tends to be thicker, and the active layer thinner, on north-facing slopes. At high elevations, permafrost can occur in temperate and even tropical regions because of the reduction in air temperature with altitude.

Vegetation affects permafrost in various ways. One [5] is by providing resistance to heat flow by conduction. By transpiration, it draws water from the soil, which is thus depleted of the heat held by the water. Furthermore, the process of evaporation (including that due to transpiration) withdraws heat from the surrounding atmosphere and from incident solar radiation. In the present context, mosses and lichens are particularly significant. Mosses are strongly hygroscopic, but can lose moisture rapidly and in large quantities. Lichens, however, have very dry surfaces at all times, even when lower layers near the soil are very wet; possibly they protect the soil against heat gain more by an insulating process than by one of evaporative cooling. It is possible, however, that rapid evaporation or diffusion exchange of water vapour from the wet basal layer to the atmosphere above the lichen may contribute to low soil temperatures and a high permafrost table. The underlying peat, formed from the accumulation of decomposed vegetation, also influences the heat transfer between the atmosphere and the soil beneath.

Snow cover influences heat transfer between the air and the ground and hence affects the distribution of permafrost. The snowfall regime and the time that snow lies on the ground are critical factors. A heavy fall of snow in the autumn and early winter will inhibit winter frost penetration and the formation of permafrost. On the other hand, a thick snow cover that persists on the ground in the spring will delay the thawing of the underlying frozen ground.

Drainage and the existence of large bodies of water greatly influence the distribution and thermal regime of permafrost. There is almost always a talik (unfrozen zone) beneath water bodies that do not freeze to the bottom. The extent of this thawed zone will vary with a large number of factors—area and depth of the water-body, water temperature, the thickness of winter ice and snow cover, the general hydrology of a lake, and the composition and history of accumulation of bottom sediments [6]. A large body of water increases the geothermal gradient around its borders [7]. For example, the influence of the ocean is of interest near its edge, where it depends upon the magnitude of the temperature difference between the land surface and ocean bottom, the thermal properties of the ground materials, past changes in climate, and the history of marine transgression and shoreline emergence. Theoretical considerations of ground temperature observations at Resolute, where permafrost is 390 metres thick at the settlement about 1·5 kilometres inland, suggest that permafrost does not exceed 30 metres beneath the ocean bottom at points more than a few thousand feet offshore.

Bare soil and rock have considerable influence on the temperature of the ground, because of their ability to reflect solar radiation. Reflectivity values in the range of 12 to 15 per cent for rock and 15 to 30 per cent for tilled soil have been observed. There will also be different evaporation rates and intakes of precipitation.

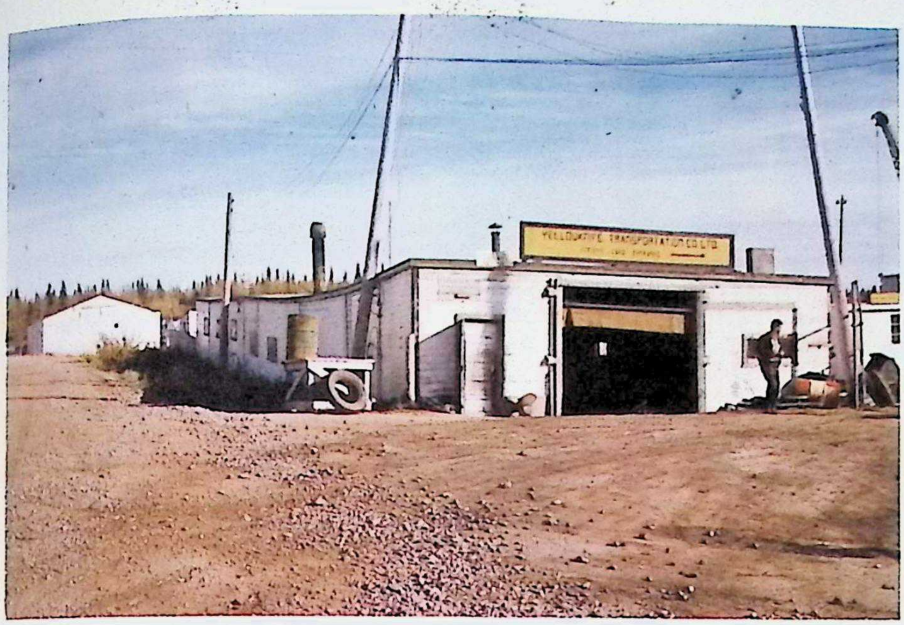
In the southern fringe of the permafrost region, most of the perennially frozen ground is confined to peat bogs. The thermal properties of the peat are to a great extent responsible for the formation of permafrost in this type of terrain; changes in the extent of permafrost are also largely dependent on changes in the thermal properties of the peat. The mechanisms which cause the formation of permafrost in these bogs are associated with variations in the heat exchange at the surface of the moss and peat. When dry, peat has a low thermal conductivity equal to that of snow (that is about 0·00017 cal/cm°C sec). Peat can absorb large quantities of water; when wet, its thermal conductivity is greatly increased. Unsaturated peat has a conductivity of about 0·0007 cal/cm°C sec, while that of saturated peat is about 0·0011 cal/cm°C sec. When frozen, its thermal conductivity is many times that of dry peat and approaches the value for ice; saturated frozen peat has a conductivity of about 0·0056 cal/cm°C sec. During the summer, a thin surface layer of dried peat having a low thermal conductivity prevents warming of the underlying soil. During the cold part of the year, the peat is saturated from the surface, and when it freezes, its thermal conductivity greatly increases. Because of this, the amount of heat transferred in winter from the ground to the atmosphere through the frozen ice-saturated peat is greater than the amount transmitted in the opposite direction in summer through the surface layer of dry peat and underlying wet peat. A considerable quantity of heat is also used during the warm period to melt the ice and to warm and evaporate the water. The net result is a loss of heat, and conditions conducive to the formation of permafrost.

This thermal mechanism does not cause the permafrost to increase in thickness indefinitely. Rather, a quasi-equilibrium is reached whereby the downward frost penetration is balanced by heat transfer from the unfrozen ground below. The thermal sensitivity of this permafrost is vividly demonstrated by its rapid thawing when the overlying moss and peat layers are removed.

What has been said above relates to the southern fringe of the permafrost region. Further north, the thermal properties of the peat and other terrain factors assume a relatively minor role and the thermal properties of the ground as a whole, together with the climate, become dominant. Variations in thermal properties such as conductivity, diffusivity, and specific heat affect the rate of permafrost accumulation. For example, the thermal conductivity of silt is about one-half that of coarse-grained soils and several times less than that of rock. Variation in thermal properties alone will not, however, necessarily result in differences in permafrost thickness within a particular area.

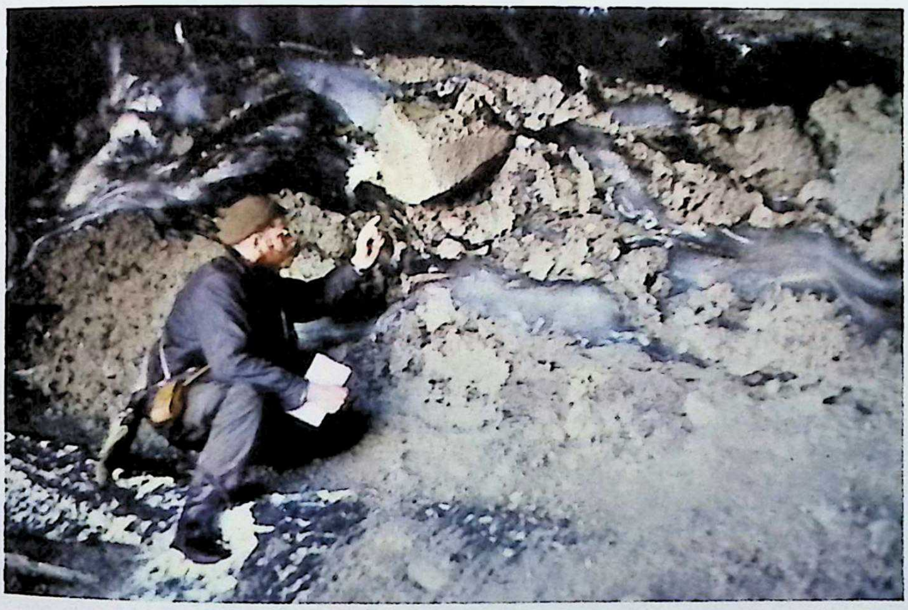
Engineering considerations

Permafrost is an important consideration in engineering work in the Far North. Although frozen soil provides excellent



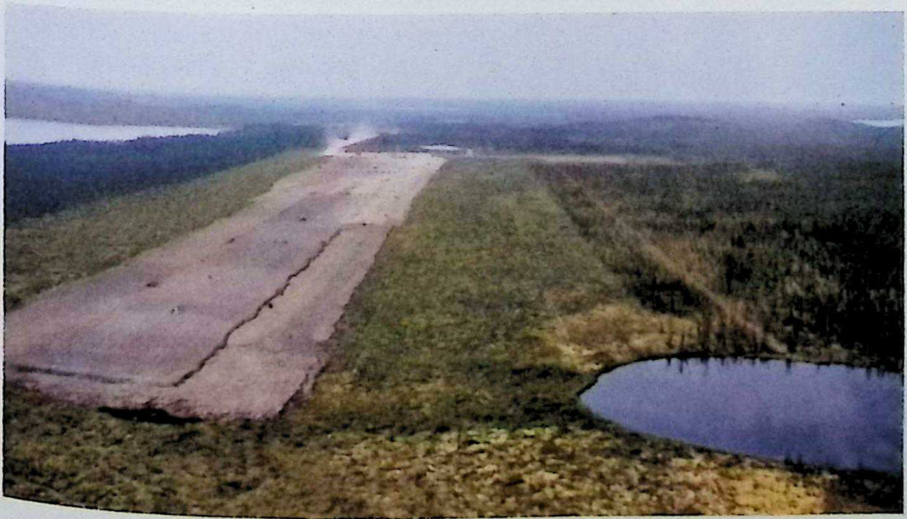
4

Figure 4 Differential settlement of warehouse on mudsill surface foundation, caused by thawing of the underlying permafrost containing large quantities of ice. Hay River, Northwest Territories, located in the discontinuous permafrost zone. September 1962.



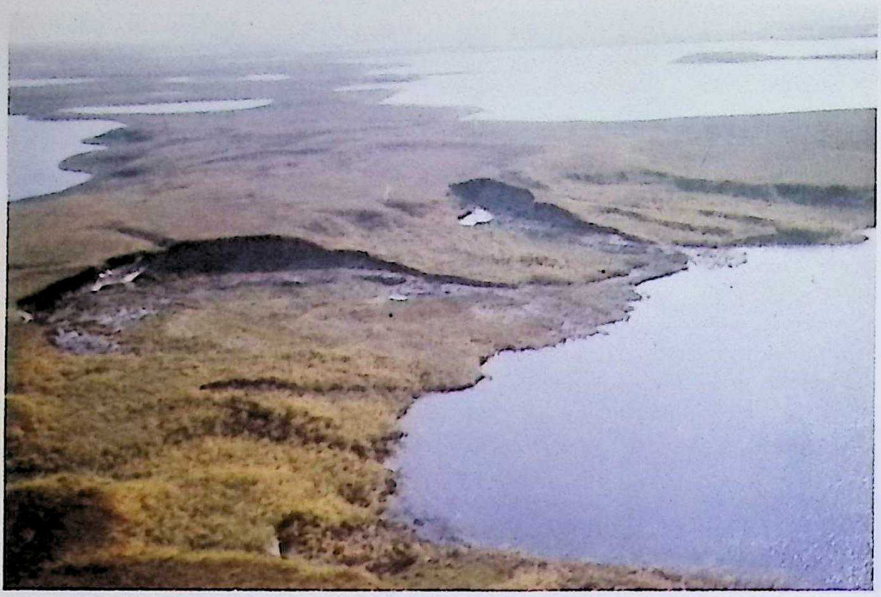
5

Figure 5 Massive ice deposits in exposure of perennially frozen stony fine-grained soil. The site is on the west side of the Mackenzie River delta, located in the continuous permafrost zone. May 1954.



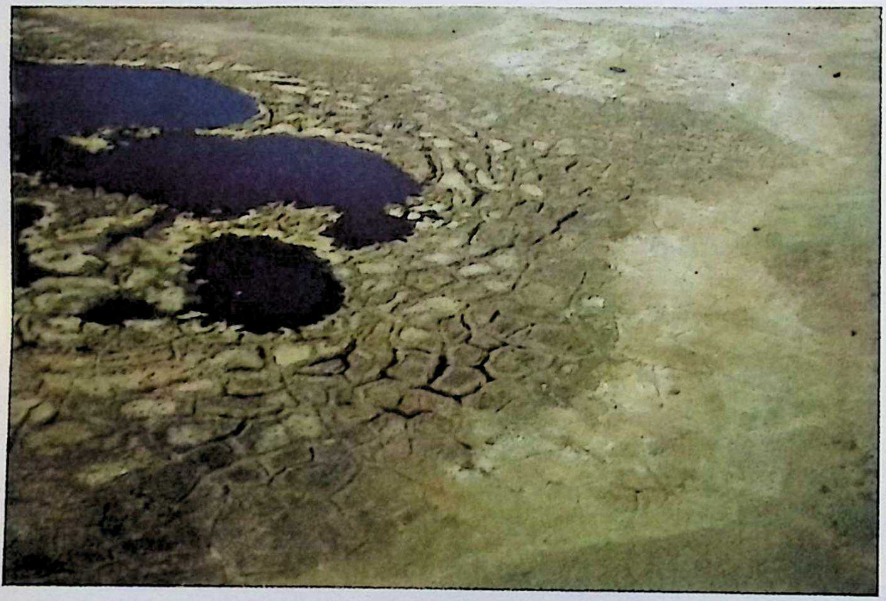
6

Figure 6 Construction of airstrip by 'insulation technique'. Placement of thick (2·5–4·5 m) crushed-rock fill on undisturbed surface vegetation cover for runway at Inuvik. July 1958.



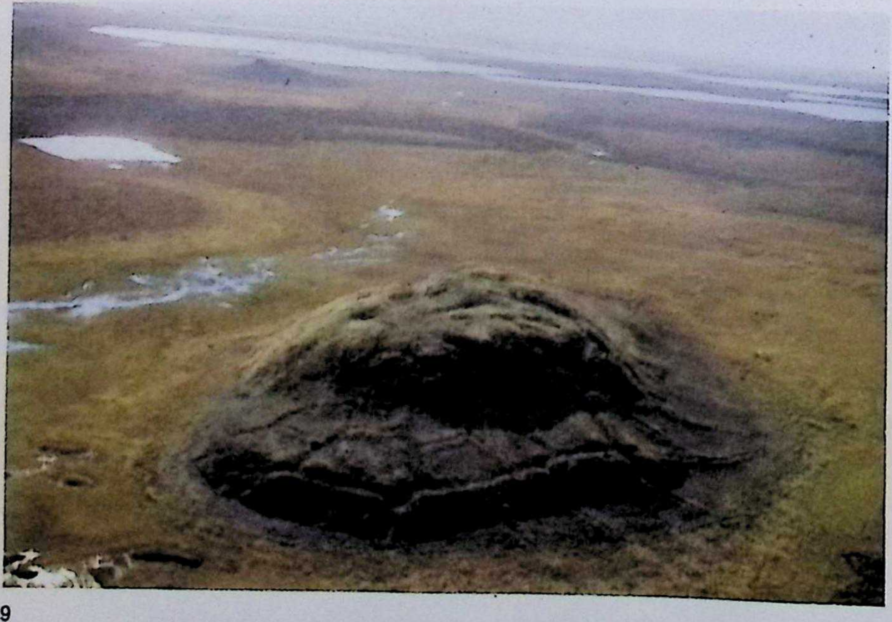
7

Figure 7 Extensive ground slumps caused by melting of massive ground ice deposits in the tundra, near Tuktoyaktuk, on the Arctic coast just east of the Mackenzie River delta, located in the continuous permafrost zone. Aerial view looking north from an altitude of about 300 m. July 1960.



8

Figure 8 Polygonal patterned ground in the tundra, located a few miles south of Tuktoyaktuk. The polygons are about 15–30 m in diameter. Aerial view from an altitude of 450 m. July 1960.



9

Figure 9 Pingo 100 km north of Inuvik. This is about 20 m high with polygonal markings on its lower slopes. Aerial view from an altitude of 60 m. September 1957.

bearing for a structure, it may when thawed lose its strength to such an extent that it will not support even light loads. Substantial settlement can occur when frozen foundations thaw, and differential movements usually result in serious damage or even complete failure of the structure (figure 4). The detrimental effect of frost action in the active layer, which alternately freezes and thaws annually, has to be considered also. Structural movements sufficient to cause damage can result from the heaving of frost-susceptible foundation soils. To ensure satisfactory structural performance, special steps, such as reducing the depth of frost penetration or replacing such soil with material in which ice accumulation will be minimal, must be taken. Three features of permafrost are significant in engineering construction:

(1) Permafrost is particularly sensitive to thermal changes. Any natural or man-made change in the environmental conditions, however slight, will greatly affect the delicate natural thermal equilibrium.

(2) Permafrost is relatively impermeable to moisture. Drainage is vital, therefore, because all movement of water occurs above the permafrost; in northern areas, surface water is conspicuous despite the generally low precipitation. If natural drainage is impeded, or proper drainage structures are not provided, construction operations can be seriously complicated by intensified frost action during the winter and accelerated thawing during the summer.

(3) The ice content of frozen ground is a most important consideration. Solid rock, gravel, and sand usually contain little or no ice at temperatures below 0°C and few difficulties are encountered in building on these materials. Most problems arise with fine-grained materials and organic materials such as peat, which usually have extremely high ice contents and are susceptible to frost action. As long as the water remains frozen in such soils, the ice binds the individual particles together to produce a material with considerable strength; when thawed, however, these soils can change to a soft slurry with little or no strength.

The ice in perennially frozen materials can occur in a number of forms: as coatings or films on small particles, stones or boulders, or as inclusions in the cavities in the soil. Such occurrences are normally observed in coarse-grained soils. Ice may also take the form of layers or lenses ranging from hairline sizes to several metres in thickness, and in this form is generally found in fine-grained deposits (figure 5). All forms can occur within the same material; in some cases, silty soils for example, the volume of ice may be as much as six times that of the soil. Some of the more spectacular ice deposits are found as large blocks or wedges in frozen ground.

Site investigations

Stereoscopic examination of air photographs, together with a knowledge of the local climate and geological history, can provide much useful information for a preliminary appraisal of an area. Surface features provide reliable indications. For example, beaded or 'button' drainage, thermokarst topography (thaw sinks), 'drunken' forest (that is, trees at various angles), solifluction lobes or terraces, frost mounds, patterned ground [8], and 'pingos' all indicate permafrost or intensive frost-active conditions (figures 7–9) [9]. The ground will contain large amounts of ice, and is potentially unsuitable for building. Selected areas, subdivided on the basis of similar terrain characteristics, are then examined in the field to check predictions made in the initial survey and to gather and correlate information on soil and permafrost conditions.

Field investigations to determine the extent and nature of permafrost are most important, particularly in the southern fringe of the permafrost region, where frozen ground is found in scattered patches and is near thawing. Variations in the depth of the permafrost table and the thickness of the permafrost over a potential site are determined during this survey. In the discontinuous zone, the depth to the permafrost table can vary from about 0·5 to 3 metres and the thickness of the permafrost from less than 1 to about 15 metres. In the continuous zone, the variation in the depth to the permafrost table is usually less, only about 0·5 to 1·5 metres, and the relatively minor variations in the great thicknesses of permafrost are of little practical significance.

Field studies must include a programme for examining and sampling subsurface materials *in situ*. This work, like the permafrost observations described above, can be carried out by probing, hand augering, or power drilling methods [10], or by the excavation of test pits. Determination of ice content is particularly important. For example, at Aklavik in the Mackenzie Delta, the top 3 metres of perennially frozen silt were found to contain 60 per cent of ice by volume. The serious engineering implications are evident, for if this frozen soil thawed there would be a settlement of 1·8 metres in the top 3 metres.

Ground temperature measurements are also an important part of these studies. This is particularly critical in the southern fringe area, where the permafrost temperatures are within one degree of zero. Further north, the ground temperatures are several degrees below zero—for example, below −5°C on the arctic coast—and consequently several degrees of warming can be tolerated without adversely affecting the strength of the foundation soil.

Design and construction

When general site conditions have been evaluated, further detailed investigations are normally required at the locations of individual structures. The results of these will indicate the approach to be taken in foundation design and the construction techniques to be used. These are usually considered under one of the following headings:

(1) Neglect of permafrost conditions.

(2) Preservation of frozen conditions for the life of the structure.

(3) Elimination of frozen condition or material before building.

(4) Thawing of frozen ground with expectation of subsequent ground settlement; foundation design takes the expected movement into account.

Perennial freezing can be neglected when engineering works are sited on well-drained coarse-grained soils or bedrock, and conventional design and construction are possible. In the zone of continuous permafrost, particularly where fine-grained soils with high ice-content are encountered, every effort must be made to preserve the frozen condition. In the discontinuous zone, it may be convenient to remove the frozen material by thawing or excavation and to replace it with well-drained material not susceptible to frost action; standard foundation designs can then be used. For some structures, in either the continuous or discontinuous zones, it may not be possible to prevent thawing of the ground during the life of the structure and settlement must therefore be anticipated and taken into account in the design.

Preservation of the frozen condition can be accomplished by either ventilation or insulation construction techniques; the former is commonly used with heated buildings. Foundations are well embedded in the permafrost, and the structure is raised above the ground surface to permit circulation of air beneath to minimize or prevent heat-flow to the frozen ground. Pile foundations placed in steamed or drilled holes have proved well suited to this method, and have been used extensively in northern Canada and elsewhere [11].

Where pile placing may be difficult, as in very stony soils, alternative foundation designs may prove more economical. Insulation to prevent thawing of the underlying frozen material may be achieved by placing a gravel blanket on the surface of the ground on which the structure is to be erected. This method is generally limited to small cheap buildings that can tolerate some movement.

For the construction of highways, railways, and airstrips [12], where the ventilation technique cannot be applied, the insulation method must be relied on (figure 6). Normally, fill methods are used throughout, and disturbance of the surface cover is kept to a minimum. Cuts through hills are avoided where possible. Proper drainage must be provided to prevent accumulation of water, which would thaw the underlying permafrost, and the formation of 'icings', which can block a road during the winter.

Excavation of frozen ground can be difficult and costly because normal excavation techniques are much less effective in permafrost. For some structures, however, it may nevertheless prove economical to excavate the frozen soil, replacing it with material not susceptible to frost action on which the foundation can be built. This method is particularly applicable in the southern fringe area. Again, adequate drainage must be provided to take care of seepage water. The procuring of large quantities of fill for road and airstrip construction presents its own problems; suitable sources must be located, cleared of vegetation, and allowed to thaw well in advance of construction operations.

Permafrost complicates the provision of water and sewer services [13]. Only limited year-round sources of water are available because many lakes and streams freeze to the bottom during the winter, and water-bearing strata are only occasionally encountered in permafrost. Normal methods of sewage disposal into the ground are generally prohibited because of the imperviousness of the permafrost. Distribution systems are generally located in insulated boxes (utilidors) on or above the ground surface because of problems resulting from excessive thawing of the frozen ground by the contents of the pipes or, conversely, freezing of pipes if they are placed in the active layer or in the permafrost.

The thawing effect of water on perennially frozen ground becomes particularly critical when dams and dykes are constructed on permafrost and large areas are covered by impounded water [14]. The rate at which thawing will take place and the depth to which thaw will penetrate beneath the water and the water-retaining structures are of prime importance in the design of their foundations. According to circumstances, underlying frozen ground can be excavated and the structure placed on bedrock; the frozen condition can be retained by natural or artificial refrigeration; or the embankment can be built up as settlement occurs when the permafrost thaws.

Conclusion

This necessarily brief review of the properties and behaviour of permafrost indicates its great significance in the development of the extensive territories in which it occurs. Not only do severe climatic conditions complicate all constructional work, but buildings may have to rest on foundations having properties quite different from those encountered elsewhere in the world. Empirical methods, assisted by growing knowledge of fundamental principles, have led to much progress but further studies are urgently needed to extend those already made, mainly in North America and the Soviet Union.

Acknowledgments

The authors wish to acknowledge many useful discussions with Dr R. F. Legget and Dr N. B. Hutcheon, respectively Director and Assistant Director of the Division of Building Research, National Research Council, Canada. This paper is a contribution of the Division of Building Research and is published with the permission of the Director.

References

[1] Muller, S. W. 'Permafrost or Permanently Frozen Ground and Related Engineering Problems'. *U.S. Army, Strategic Engng. Study* 62, 1943.
[2] Hopkins, D. M., Karlstrom, T. N. V., *et al. Geol. Surv. Prof. Paper* 264-F, Washington, 1955.
[3] Shvetsov, P. F. (Editor). Основы Геокриологии (мёрзлотоведение) (Fundamentals of Geocryology (permafrost)). I. Academy of Sciences of the U.S.S.R., Moscow, 1959.
[4] Brown, R. J. E. *Arctic*, **13**, 163, 1960.
[5] Legget, R. F., Dickens, H. B., and Brown, R. J. E. in 'Geology of the Arctic', Vol. II, p. 956, University of Toronto Press, Toronto, 1961.
[6] Johnston, G. H. and Brown, R. J. E. *Nature, Lond.*, **192**, 251, 1961.
[7] Lachenbruch, A. H. *Bull. geol. Soc. Amer.*, **68**, 1515, 1957.
[8] Washburn, A. L. *Ibid.*, 67, 823, 1956.
[9] Sjörs, H. *Endeavour*, **20**, 217, 1961.
[10] Hvorslev, M. J., Goode, T. B. Boyd, W. K., and Lange, G. R. 'Core Drilling in Frozen Ground'. Seventh Annual Drilling Symposium. University of Minnesota. 1957.
[11] Pihlainen, J. A. *Proc. Amer. Soc. civ. Engrs,* (*Soil Mech. Div.*), **85**, SM1, 1, 1959.
[12] Linell, K. A. *Ibid.*, (*Air Transp. Div.*), **83**, AT 1, 1, 1957.
[13] Dickens, H. B. *Polar Rec.*, **9**, 421, 1959.
[14] Lewin, J. D. *Publ. Wks, N.Y.*, **79**, No. 5, 22; No. 6, 33; No. 7, 57, 1948.

Magnetohydrodynamics

W. B. Thompson

Magnetohydrodynamics may be said to stem from Michael Faraday's attempt to measure the electrical potential across the Thames due to the river's flow through the Earth's magnetic field. Only comparatively recently, however, has the effect been extensively studied. The results have been used to explain such diverse phenomena as the generation of cosmic rays and the maintenance of the Earth's magnetism. Among potential applications that have been considered are the propulsion of space ships, the direct conversion of heat to electricity, and controlled thermonuclear fusion.

The two basic principles of magnetohydrodynamics (MHD) are extremely simple and familiar. One is that a force is exerted on a conductor carrying an electric current across a magnetic field; it is this force that is employed in the electric motor. The second is the production of an electric current in a conductor made to move across a magnetic field; this is the principle of the dynamo. Clearly, if an electrically conducting fluid carries a current across a magnetic field, each element of the fluid will experience a force. This force (F) is proportional to the product of the current density (j) and the component of magnetic field normal to the current (B), and directed at right angles to both. Mathematically this is expressed as

$$F = j \times B.$$

Such a force will obviously affect the flow of the fluid. At the same time, if the fluid moves across lines of force, an electromotive force will appear proportional to the rate at which flux is cut, and this in turn may give rise to an electric current. Thus motion gives rise to an electric current that modifies the magnetic field and at the same time interacts with it to produce a force on the fluid. Although the individual processes in a conducting fluid are simple in themselves, they are related in such an intimate way that the total phenomena may be quite unexpected.

Channel flow

The simplest MHD phenomena are those that occur when a fluid runs through a channel crossing a magnetic field. When this happens, we expect, by Faraday's law of induction, to find a potential, V, across the channel equal to the rate at which flux is being cut.

$$V = 300(v/c)BL = 10^{-8}BLv \text{ volts}$$

where v is the flow velocity in cm/sec; L the width (or average width) of the channel in centimetres; B the field strength in gauss; and c the velocity of light.

Faraday argued that if this potential could be measured, the mean rate of flow of fluid through the channel could be determined. One January night in 1832, he tried to measure the potential developed across the Thames at Westminster Bridge, London, as a result of the river flowing through the Earth's magnetic field. The expected potential was rather small, $\sim 0{\cdot}05$ volts, but this first measurement produced an answer not only of the right sign, but of approximately the expected magnitude. To confirm this result, he repeated the measurement the next day and suffered a disappointment, now only too familiar in magnetohydrodynamic experiments, by finding a voltage of the wrong sign. However, since this first inconclusive experiment, the electromagnetic flowmeter, which involves exactly this principle, has been developed as a useful instrument (figure 1). In one form, it has been employed to measure the flow of blood through capillaries; in another, to measure the flow of the ocean through the English Channel. Nuclear engineers use another version to measure the flow of liquid metals through heat exchangers.

If electric currents are permitted to pass through the fluid, then forces will act on it. In a strong enough magnetic field, the motion of the fluid may be violently altered. An early demonstration of this effect was given by J. Hartmann, who allowed mercury to flow through a narrow glass tube between the poles of an electromagnet. He found that the pressure required to produce a given flow increased with the field strength. He attributed this to a non-uniformity of velocity across the tube, which would cause currents to flow internally. The electromagnetic forces redistributed the flow across the tube, making it more nearly uniform and hence increasing the viscous drag on the walls. A second phenomenon he observed was delay in the onset of turbulence in a magnetic field. His observations of pressure and flow-speed showed a distinct break at a high rate of flow, the additional pressure needed to produce a further increase in flow suddenly becoming much greater. This break is usually attributed to the onset of turbulent flow, and Hartmann found that the flow-speed at which the break occurred increased with increase in the magnetic field (figure 2). The effect of a strong magnetic field on the forces required to keep a fluid in motion will be clearly felt if one stirs a beaker of mercury placed between the poles of a magnet. The normally labile fluid feels as viscous as treacle;

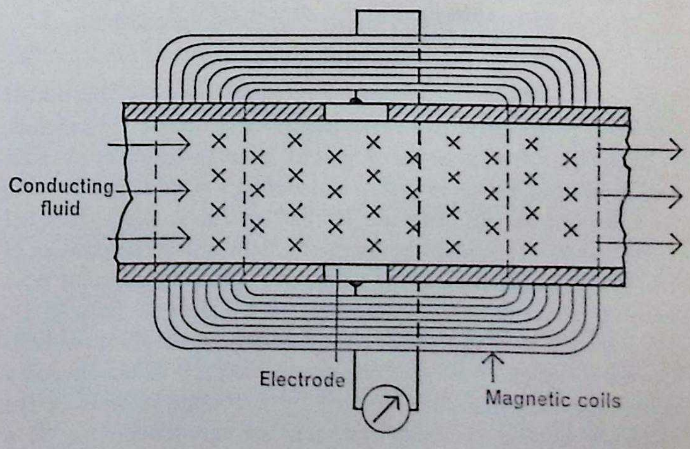


Figure 1 Electromagnetic flowmeter.

W. B. Thompson, *M.A., Ph.D.*
Was born in Belfast, Ireland, in 1922 and educated at the Universities of British Columbia and Toronto. In 1950, he joined the Theoretical Physics Division of the United Kingdom Atomic Energy Authority at Harwell, and began theoretical studies relating to the controlled release of thermonuclear power; this has since remained his main field of interest. In 1961, he became Head of the Theory Division of the U.K.A.E.A.'s Culham Laboratory, and last year he was appointed Professor of Theoretical Plasma Physics at Oxford University. In 1961–2 he was a Research Associate at Princeton University and Visiting Professor at the University of California, La Jolla.

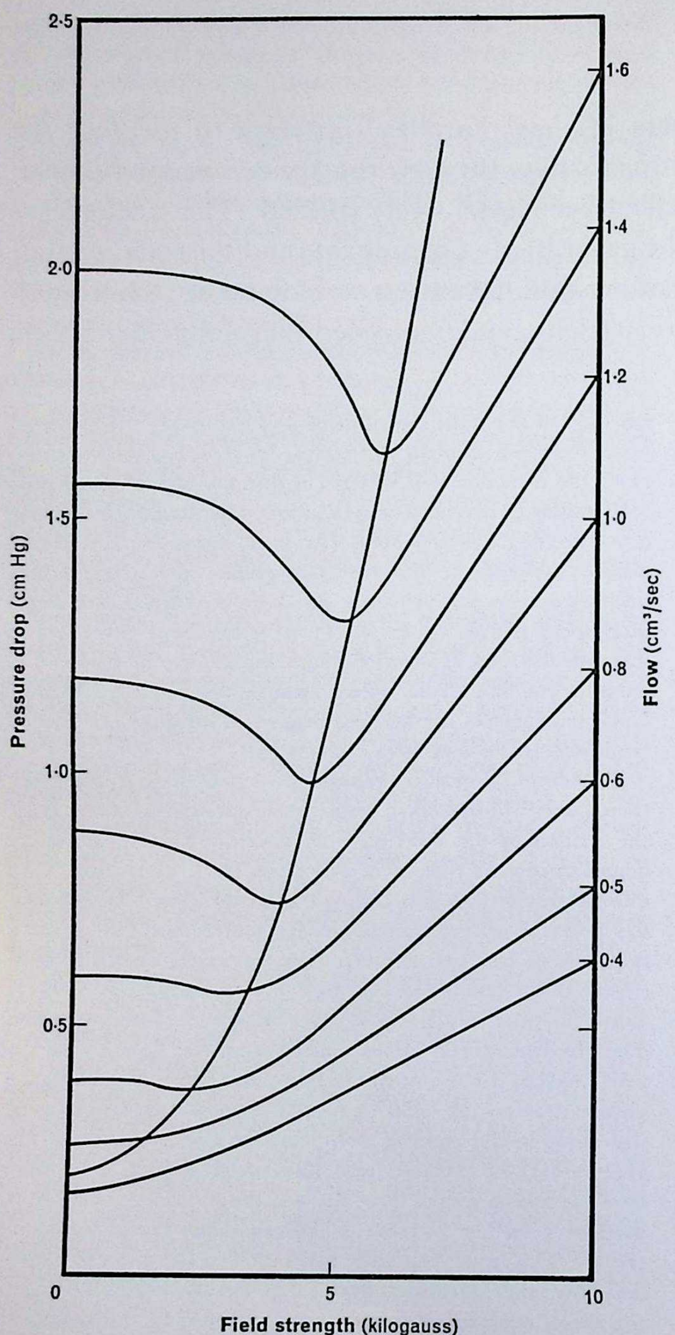


Figure 2 Flow of mercury across a magnetic field. Curves showing pressure as a function of magnetic field strength at fixed rate of flow through a rectangular tube 1·82 mm × 2·15 mm in cross section, 140 m in length. Flow is turbulent to left of the parabolic curve, which demonstrates the inhibition of turbulence by a magnetic field (*K. danske vidensk. Selsk. Nat-fys. Medd.*, **15**, 6, 1957).

indeed, in a fluid of moderate electrical resistivity this added viscosity is the principal consequence of magnetic interaction.

If the fluid-flow is uniform, and all boundaries of the channel are insulators, no current flows, although a potential difference appears across the channel. If this voltage is short-circuited currents can flow, and the magnetic field resists the passage of the fluid. Thus, if mercury flows through an insulating duct passing across a vertical magnetic field it does so quite freely. However, if two electrodes on either side of the duct are connected by an external wire, the flow is violently disturbed and a hydraulic discontinuity, marked by a change in level of the mercury surface, extends between the two electrodes. If a copper sheet is placed on the floor of the duct, it in turn short-circuits the electromotive force due to motion. The mercury over the copper is reduced almost to rest, and the remainder of the flow passes around it, forming a wake, just as if it were a solid object. These phenomena have been spectacularly demonstrated in experiments conducted by R. A. Alpher and H. Hurwitz.

That a current will flow between electrodes placed on either side of a channel carrying a conducting fluid across a magnetic field has what may prove to be an important practical application. If the electrodes are connected through an external circuit, the process provides a method of converting the energy of fluid flow directly into electrical energy. The conducting fluid under consideration in this context is an ionized gas produced in a flame. At high enough temperatures, and modest pressures, the violence of collisions between the molecules in a gas becomes great enough to knock electrons off them. Since the electrons so released are free to respond to electric fields, the ionized gas acts as an electrical conductor, and is known as a plasma. In normal flames, the degree of ionization is very small, but the addition of small amounts of a salt of an alkali metal such as potassium or caesium, from which electrons are easily detached (a process known as 'seeding') increases the degree of ionization until the flame gases are quite good electrical conductors. By passing such a plasma between conducting electrodes and across a magnetic field, it should be possible to generate reasonable amounts of electric power by a device that has no moving parts (figure 3). Such a process would form a useful addition to those conventionally used in power-stations. The turbines in these are limited to maximum working temperatures of around 500°C so that energy can be extracted only after the flame has been allowed to cool. Since flames have temperatures of 1000–1500°C, a turbine necessarily operates at only a fraction of its potential thermodynamic efficiency. Even a modest efficiency in direct conversion could produce a significant increase in the efficiency of power generation, and be enormously profitable. Electricity is generated in such a system at a fairly low potential; for example, across a 10 cm gap in a field of 10 kgauss, a gas consisting of combustion products produces an electromotive force of approximately 100 volts (figures 3 and 4).

There are several possible methods of extracting energy from gas flowing through a channel. The flow may be initially subsonic, in which case the gas pressure acts against the pressure of the magnetic field; consequently the gas expands and cools. On the other hand, the gas may be expanded before reaching the magnetic field and be made to encounter the magnetic field at supersonic speed. In this case, the kinetic energy of the gas is the source of the electrical energy, and the

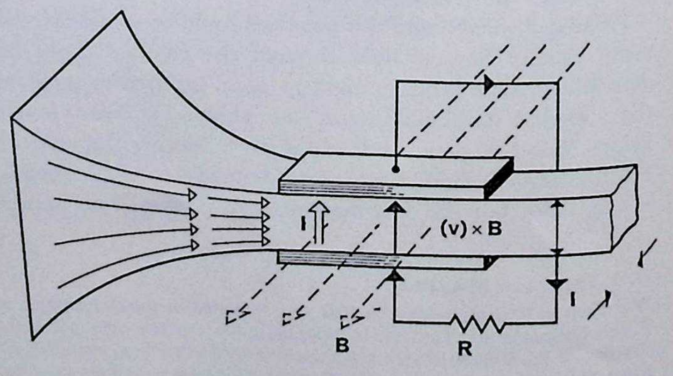


Figure 3 Direct conversion of kinetic to electrical energy. Flow of conducting fluid across field **B** induces a Lorentz force $(v) \times \mathbf{B}$ on charges and current *I* in load resistor *R*. Pressure increases, but total energy of fluid decreases.

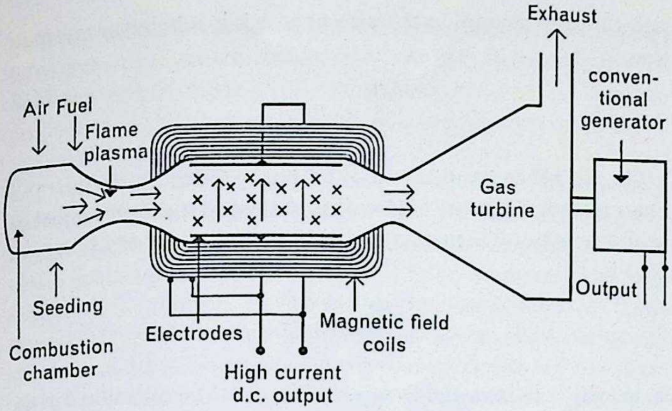


Figure 4 Schematic illustration of possible direct conversion generator.

gas is compressed, and actually heated, in passing through the magnetic field. The increase in temperature results from the kinetic energy of flow as, of course, the total energy in the gas is reduced. Experiments of this second kind have been conducted by H. J. Pain of Imperial College, London, in shock tubes; quite high fractions (~30 per cent) of the energy in the gas have been extracted as electrical energy. Of course, the duration of flow in a shock tube is only a few microseconds and the total energy extracted is modest, although the power level is high—several megawatts. It is particularly satisfactory that these experimental results are in good agreement with a very simple theory of the process, which, by considering the budget of momentum and energy for one-dimensional flow, predicts the conditions necessary for maximum power extraction.

More spectacular experiments have been made using continuously burning flames in seeded gases, and 5 megawatts have been extracted continuously in an experiment conducted at the AVCO Laboratory in the United States by a group now developing a 50 megawatt generator.

In 1879, Hall discovered that if current flows through a thin metal plate placed in a magnetic field whose lines of force are perpendicular to the direction of flow, a potential difference exists between opposite points in the plate at right-angles to the direction of the current. In ionized gases, the Hall effect is large. As an electric current does not flow in the direction of the applied field, but at some angle to it, it follows that when the gas flows through a duct transverse to a magnetic field, the current does not necessarily flow across the duct. If it does, it is not necessary for a potential to appear across the duct. Consequently, the electrodes may take the form of equipotential rings about the channel. Flame experiments with Hall-effect generators conducted by J. Wright of the Central Electricity Generating Board have extracted modest power at low efficiency from simple seeded flames.

The major problems awaiting solution in the direct conversion of heat to electricity have little to do with MHD itself, but are almost entirely concerned with the production of ionization or with the prevention of wall and electrode corrosion. Indeed, the large AVCO experiments made no attempt to overcome this latter problem, but were conducted in a channel with wooden walls which burned away during the experiment. The use of seeded gases is expensive, because if they pass into the atmosphere, the seeding material is lost. If they pass out through a turbine, corrosion problems are aggravated. In a closed system, the seeding material could be recovered, but the working fluid would have to acquire its energy through some high-temperature heat exchanger. This might be feasible in some future ultra-high-temperature nuclear reactor, but in present day contexts the project appears unattractive. Direct conversion research represents, however, one possible application of MHD that has reached the level of realistic engineering at which economic considerations play a vital role.

The reverse of the above processes also has interesting possibilities. It is possible, by using an external generator, to reverse the direction of flow of the current between the electrodes and thus to add energy to the fluid instead of extracting it. A simple application is to the pumping of liquid metals, which is particularly important in handling the alkali metals used for heat-transfer in nuclear reactors. A major problem here is to maintain efficiency at low flow-speeds. The power supplied to the system is the product VI of the current flowing through the metal and the voltage drop across it, while the power used for driving the flow is V_bI, where V_b is the back electromotive force.

$$V_b = \left(\frac{v \times B}{c}\right) L = 10^{-8} vBL \text{ volts}$$

v being the mean flow velocity in cm/sec, and L the channel width in cm. For modest flows, and attainable magnetic fields, this back e.m.f. is necessarily rather small. Consequently, power must be delivered at low voltage; this in turn requires that the power supply should be of extremely low impedance, so that its internal energy dissipation is kept low.

A more fanciful version of the electromagnetic pump makes use of the increase of efficiency with increase in speed of flow, and finds its possible application where extremely high flow-speed is required, for example in the propulsion of interplanetary space-ships. In order to carry out an interplanetary journey in a reasonable time, a vehicle must move with an extremely high velocity, of the order of 10^6–10^7 cm/sec. To attain such a high speed, a large impulse must be given to the vehicle, and in outer space, this impulse must be provided by hurling matter backwards. The total impulse, and hence the ultimate velocity attainable by a vessel of given mass, can be increased either by increasing the amount of matter thrown off or by increasing the speed with which it is hurled. If the former method is used, however, the extra propellant must be carried with the vehicle for a least part of the journey and thus must itself be accelerated. Hence, to get a modest increase in terminal velocity an enormous additional mass of propellant is required. However, an increase in exhaust velocity involves no such disadvantage, and the terminal velocity, V_f, is related to the exhaust velocity, V_e, by the equation

$$V_f = V_e \ln (M_i / M_f)$$

where M_i is the initial mass and M_f the final mass, that is, when all fuel is burnt.

In practice, however, increasing V_e is a formidable problem. The speeds attained by the molecules in a chemical reaction, approximately 10^5cm/sec, are a good deal too low, but if electrical energy is added to an ionized gas, then the attainable exhaust-gas speeds are virtually unlimited. Of course, an increase in exhaust velocity calls for an increase in the energy per unit mass supplied to the propellant and if the energy is released by burning the propellant, as in a rocket, there is little point in accelerating a fraction of the fuel at the expense of the rest. If, however, some non-chemical source of energy is tapped—such as solar or nuclear energy—then this energy may be used as electrical power and the propellant accelerated by magnetohydrodynamic means to velocities much greater than those obtainable chemically. This would result in an enormous reduction in the initial mass of a vehicle, but would have to be paid for in a severe reduction in thrust, thus limiting high exhaust velocity drives to interplanetary space, where other forces acting on the vehicle are negligible.

To obtain the enormous velocities required while using only reasonable amounts of energy, the propellant must have very low density, and, as in direct conversion, one must use a diffuse ionized gas as the working fluid.

To avoid using separate circuits to produce the driving field B and the driving current j, the current j in the gas may be allowed to interact with the field it itself produces. This is easily done by releasing the gas in successive puffs between two charged metal rails, or into a charged coaxial system (figure 5). The ionized gas then forms the closing link in the electrical circuit, and moves, according to Lenz's law, in such a way as to increase the inductance. The efficiency of the system can then be expressed in terms of the ratio of change in inductance to final inductance, resistive losses being neglected. A major problem, therefore, is to produce a current supply with extremely low inductance. Experiments carried out with plasma guns, as these devices are called, have in fact produced bursts of ionized gas moving with speeds of $\sim 10^7$ cm/sec. However, before these laboratory experiments produce a practical interplanetary drive many problems, such as reduction of weight and size and, above all, increase in reliability, must be solved. On an unmanned space-vehicle, an engine must run unattended for several years, and reliability is particularly important. On the other hand, the problems seem intrinsically straightforward, and no doubt MHD drive will play an important role in future space exploration.

'Ideal' magnetohydrodynamics

To investigate MHD behaviour in geometrically more complicated systems, a slightly more sophisticated language is required than that which proved adequate for one-dimensional channel flow. Just as in normal hydrodynamics, various idealizations are usually made about the properties of the fluid in order to get an intelligible model of its flow. For example, in hydrodynamics one idealizes the fluid as incompressible for some purposes, or considers the flow as adiabatic for others, only very rarely taking into account the full complexities of its actual physical properties. A customary additional idealization in MHD is to consider the fluid as a perfect conductor.

In a perfectly conducting fluid, the changes produced in the magnetic field are directly related to the motion of the fluid, for the magnetic flux threading any closed loop sharing the motion must remain a constant. If it did not, there would appear, by Faraday's induction law, an electromotive force about that loop. But a perfectly conducting fluid cannot sustain an electric field, and consequently the electromotive force must vanish.

To say that a fluid cannot sustain an electric field is a somewhat ambiguous statement, since the electric field in a moving system must be defined with some care. The statement means that the force acting on a unit charge sharing the local velocity v of the fluid must vanish. If the electric field E, however, is taken as that measured by some device at rest, then the total force acting on a unit charge is, in addition to that due to the electric field, the Lorentz force $v \times B$. The electric field E^*, as seen by the fluid, is just the sum of these forces

$$E^* = E + v \times B$$

It is this quantity that must vanish in a perfect conductor, and whose vanishing ensures that the magnetic flux is conserved.

If the magnetic field is not uniform in the fluid, but changes over some distance L, then the rate of change of field at a point is proportional to the rate at which it is being swept past, that is,

$$\frac{dB}{dt} \propto vB/L$$

On the other hand, if the fluid has a finite resistivity, η, a non-uniform magnetic field will tend toward uniformity, just as would a non-uniform temperature, and the rate of change of field is

$$\frac{dB}{dt} \propto \eta B/L^2$$

from which follows the familiar 'skin depth' expression for the time $t \sim L^2/\eta$ needed for a field to penetrate a conductor.

If the change in the magnetic field is due almost entirely to convection, that is, if $vB/L \gg \eta B/L^2$, the fluid will behave almost as a perfect conductor: that is, if the 'magnetic Reynolds' number' $R_M = vL/\eta \gg 1$, the perfect conductivity idealization is useful. This is large for large volumes of rapidly moving material having low resistance. The large values of L in astrophysical and geophysical phenomena mean that idealized MHD is an appropriate language for discussing them. For example, Lamb's calculation of the penetration time for the Earth's magnetic field (about 50 000 years) suggests that the secular variation of the Earth's field, which has a scale length representing a considerable fraction of the Earth's radius, and a time scale of about 100 years, represents a MHD phenomenon in the Earth's core. Similarly, the magnetic fields in sunspots, which have dimensions of about 10^4 km and durations of months, appear to be a MHD phenomenon in the surface layers of the Sun.

At the other end of the scale, the high velocities and short durations associated with many laboratory experiments in plasma physics—such as the compression of ionized gases by rapidly rising magnetic fields, where velocities are about 10^6 cm/sec, and the duration of the experiment about 10 μsec—suggest that the neglect of resistive diffusion of the magnetic fields may be justified even for small-scale lengths around 10 cm.

To discuss idealized MHD, some familiar ideas originally introduced by Faraday are extremely useful; such are the concepts of tubes of flux and lines of magnetic force. Since flux is conserved, one may think of the fluid as glued to the lines of force, and responding to the forces between flux tubes. Faraday showed that the configuration of a static magnetic field could be understood by considering that neighbouring lines of force repel one another with a force $B^2/8\pi$, and that in each line there is a tension of similar magnitude. As a result, the stress in the fluid includes, in addition to the hydrostatic

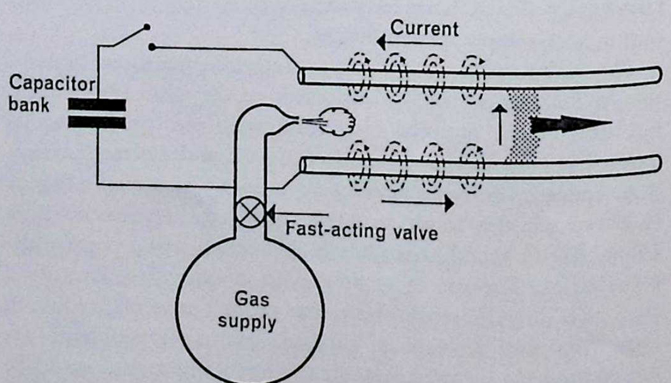


Figure 5 Principle of the rail gun. A burst of gas released by a fast-acting valve is ionized by condenser discharge and driven down the rail by the self field.

pressure, p, the anisotropic stress of the magnetic field. The force per unit volume responsible for the acceleration of a fluid element becomes

$$-\nabla(p+B^2/8\pi)-(B\,.\,\nabla)B/4\pi$$

Of these, the first magnetic term represents the repulsion between neighbouring flux tubes, while the second makes two contributions; it cancels out the component of repulsive stress along the magnetic field lines and, if the lines are curved with radius of curvature R, introduces the contribution of the tension $B^2/4\pi R$ to the accelerating force. The magnetic stresses need not be small, for the energy density of the magnetic field ($B^2/8\pi$) is quite large for reasonable fields. If $B \simeq 5000$ gauss, this is approximately 10^6 erg/cm^3, the equivalent of one atmosphere pressure. When acting on diffuse ionized gases, the magnetic field may completely dominate the forces, the gas doing little dynamically except to provide inertia.

The effect of these added stresses will be felt if the fluid is disturbed. In particular, the possible response of the fluid to a small disturbance becomes much more varied than in normal hydrodynamics. A normal ideal fluid will transmit only one class of oscillation, the compressive sound waves, and any non-compressive distortion of the fluid is localized. In MHD, the magnetic field confers a rigidity on the fluid, and if, for example, the field lines are locally twisted in such a way that the fluid is not compressed, the twist will be propagated along the field lines as an Alfvén wave with a velocity, C_A, the Alfvén speed, determined by the interplay between the magnetic stress and the inertia of the fluid of density ρ.

$$C_A=\sqrt{\left(\frac{B^2}{4\pi\rho}\right)}$$

On the other hand, a wave in which the fluid is compressed only in the direction of the magnetic field will propagate with the normal velocity of sound $C_S=\sqrt{(\gamma p/\rho)}$.

If compression of the fluid and distortion of the field both occur, there are two possible ways in which they may be coupled. The magnetic and hydrostatic stress may cooperate, giving a wave speed that varies between C_S along lines of orce and $(C_A{}^2+C_S{}^2)^{\frac{1}{2}}$ across them for strong magnetic fields. Alternatively, they may oppose each other, giving a wave speed which varies between zero across, and C_S along, lines of force. In a strong magnetic field, this latter slow wave reduces to a sound wave, but one which is channelled along the magnetic lines of force, the field resisting its propagation in any other direction.

A peculiar property shared by the slow wave and the twisting Alfvén wave, is that they propagate only along the magnetic field lines. Hence, even if initiated by a point source, they undergo no geometrical attenuation corresponding to the inverse square attenuation of a normal sound wave.

Of these possible phase velocities, only the Alfvén velocity has been clearly demonstrated experimentally, for only in an ionized gas can experiments be made, and most devices produce plasmas that are irregular, and too 'noisy' for disturbances in the gas pressure to be detected. The Alfvén wave which distorts the magnetic field may be fairly well identified, for the magnetic field may be disturbed electrically, and the propagating disturbance detected electrically at a different place (figure 6). Indeed, the waveguide modes for these oscillations in a plasma cylinder have been both calculated and experimentally investigated, with very satisfactory results. The slow wave has been invoked by Lighthill to explain the erratic behaviour of the radar cross section of rockets passing through the ionosphere. As an object goes by, Lighthill envisages these slow waves being propagated along the Earth's magnetic field lines, increasing the electron density and producing a radar-reflecting wake whose form depends on the orientation of the trajectory, the radar beam, and the Earth's field. The effect of these peculiar oscillations on the production of wakes is entertaining. In MHD, in contrast to normal hydrodynamics, a wake need not spread out symmetrically about a moving object but goes in a direction determined by the magnetic field, sometimes even opening ahead of the moving source.

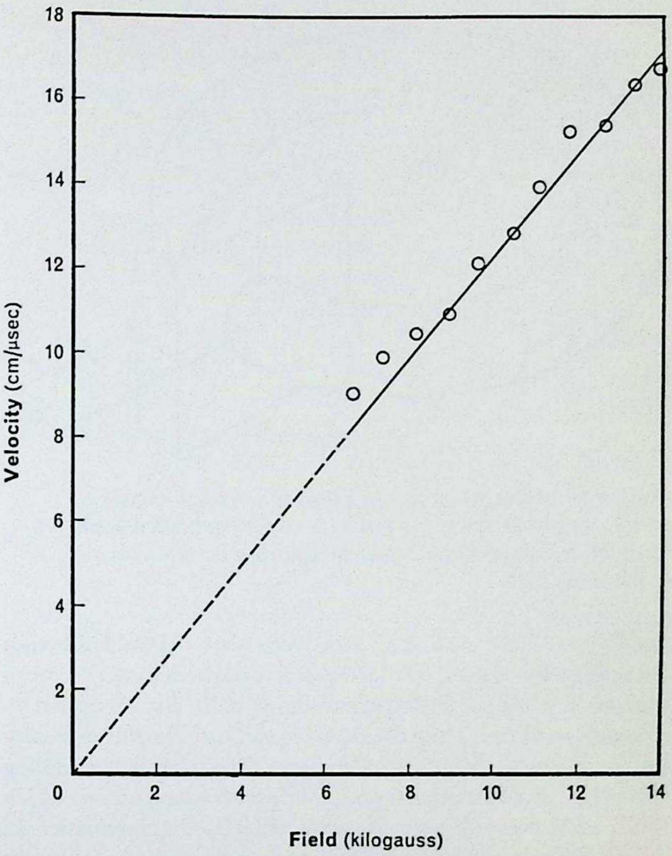


Figure 6 Velocity of Alfvén waves. Measured speed of magnetically produced Alfvén wave in a low-pressure argon discharge in a strong magnetic field. The straight line is the theoretical value (Hardcastle and Jephcott, A.E.R.E. report R3, 109).

Wakes are affected by a magnetic field, as also is boundary-layer flow. In normal hydrodynamics, a boundary layer is confined to fluid that has been very near to the disturbing boundary, since the vorticity produced by the boundary can be convected only with the fluid. In MHD, however, the Alfvén waves carry pure vorticity along magnetic field lines, and there is no reason why the boundary layer should remain attached.

The propagation of Alfvén waves through the ionosphere has been invoked by a number of investigators to explain slow oscillations of the Earth's field having periods ranging from small fractions of a second to over a minute. Although the data are suggestive, no clear demonstration has been given that this is in fact the correct explanation.

Magnetic confinement

An interesting possibility suggested by the nature of magnetic stresses acting on a conducting fluid is that such a fluid could be held in place against gravity, or against its own pressure, by magnetic forces. One application of this idea is made in the levitating furnace. In this device, a small piece of metal is placed in a rapidly oscillating magnetic field produced by coils so situated that the metal lies in a hollow in the field. The

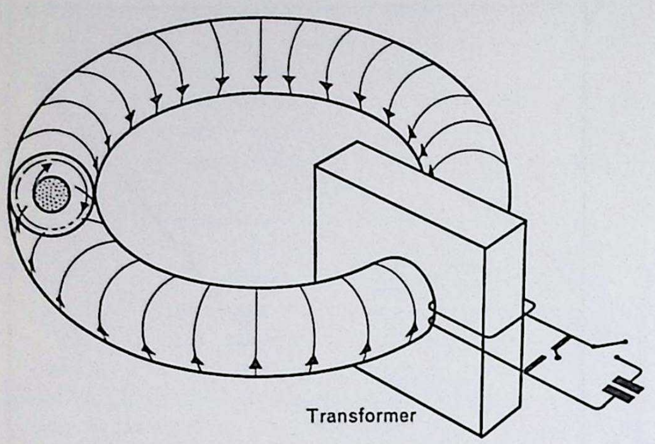


Figure 7 The toroidal pinch. External windings supply a stabilizing field along the torus. A capacitor bank discharged through a pulsed transformer produces a current and a collapsing field.

oscillations have such high frequency that the field does not penetrate the metal, while the fluctuations about the mean magnetic pressure $B^2/8\pi$ occur too rapidly for the metal to respond to them. Thus the metal is supported against gravity by the magnetic field and at the same time is heated to melting point by the skin currents on its surface. Such a furnace which both levitates and heats a metal permits the production of alloys from components too reactive to be melted in contact with the walls of a material furnace and represents a first essay in non-material confinement.

This possibility of confining material in a magnetic container, isolated from any material contact, suggests one of the most exciting applications of MHD. If the pressure of an ionized gas could be balanced by magnetic stresses and the gas heated by changing electromagnetic fields, then—since it can lose heat only by radiation, a rather inefficient process when the gas is completely ionized—it should be possible to reach temperatures of 10^7 degrees, exceeding the central temperature of the Sun. At these elevated temperatures, the collisions between nuclei are so violent that nuclear, as opposed to chemical, reactions are quite probable. If the heavy isotopes of hydrogen are used, the predominant reactions would lead to the formation of helium and the release of large amounts of energy. Since heavy hydrogen is fairly common, and the fusion process produces none of the dangerous fission products that arise in the burning of uranium or plutonium, the possibility of tapping this fusion energy in a controlled manner would be of the utmost practical importance.

These studies began by investigating extremely simple geometries, such as the 'pinch' (figure 7). Basically, this is a simple column of plasma; a current flowing along its axis produces a magnetic field about the axis that confines the plasma to a cylindrical column. In order to act as a confining system, the column may be bent into a circle, and the current produced by induction, so that the plasma acts as the secondary coil of a transformer. A second simple geometry may be obtained by interchanging the role of current and magnetic field, that is, by applying an axial magnetic field to the plasma column, and inducing azimuthal currents about it. When an attempt is made to bend this column into a circle, a difficulty arises, because the magnetic force $j\times B$ then becomes rotational, and cannot be balanced by a pressure gradient. This may be avoided by twisting the column not into a circle, but into a figure of eight—the original stellerator geometry (figure 8). Alternatively, the column may be made of finite length by pinching it off by means of a strongly increased magnetic field at the two ends, a geometry exploited in the thetatron or 'mirror'.

These early experiments all produced rather unsatisfactory results; either the plasma was confined for an unexpectedly short time, or the power losses were unexpectedly high. When the plasma was examined with high-speed cameras, the reason for these anomalies became clear. The plasma shape very rapidly became violently distorted; large ripples appeared on its surface, or the column wriggled into some irregular helix and drifted away from the centre of the containing vessel, or it broke up into isolated blobs, in one way or another making contact with the cold walls of the vessel. These processes all represented the development of instabilities in the original equilibrium, and the study and control of these instabilities have formed a major part of research into controlled thermonuclear reaction in recent years.

It is, in principle, fairly easy to make a theoretical study of the stability of a static MHD equilibrium by investigating the way in which the energy changes when the system is slightly disturbed. The method is simply illustrated by a bead sliding on a wire that is curved in a vertical plane. Such a bead can rest wherever the wire is locally horizontal. If the horizontal point is at the bottom of a curve, the bead's potential energy is increased by any small displacement, and it is, of course, at a stable point. On the other hand, if it is at the top of a hump, any displacement will decrease the potential energy, and the bead will slide away.

The potential energy E of a MHD equilibrium is the sum of the magnetic energy density $B^2/8\pi$ and the thermal energy $p/(\gamma-1)$ for each volume element in the fluid, and, if the fluid is in a gravitational field, the gravitational energy $\rho\varphi$: then

$$E=\int dv[B^2/8\pi+p/(\gamma-1)+\rho\varphi]$$

[φ is the gravitational potential]

If the configuration is changed, three things may happen. The convection of magnetic flux with the fluid may produce a change in the magnetic energy; compression or expansion of the fluid may alter the thermal energy; and a raising or lowering of the fluid in the gravitational field may alter the potential energy.

As an example of the first sort of distortion, consider a simple pinch in which the fluid is confined by a surface current. If a ripple is allowed to form about the cylinder—some fluid moving out and some magnetic flux moving in, so that the volume of the fluid is unchanged—the magnetic flux lines will have shortened, and the magnetic energy have become reduced, so that the system is unstable. If, on the other hand, the positions of field and plasma are reversed, then the surface is stable against ripples, since an exchange of fluid and field would result in a lengthening of lines of force and an increase in magnetic energy. This would happen, for example, if a current

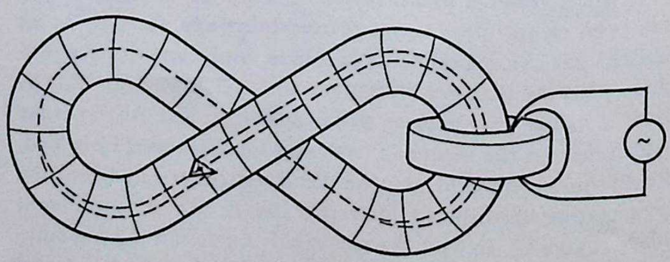


Figure 8 The stellarator. A d.c. confining field is produced by external windings on the tube, which to satisfy equilibrium requirements must be deformed into a figure of eight. The transformer produces an axial current for heating the plasma.

flowing through a rigid conducting cylinder generated an azimuthal magnetic field. This field, by producing surface currents in a conducting fluid, would keep the fluid out of a cylindrical cavity.

Experiments by S. A. Colgate provide some direct evidence that this field-curvature effect does have the expected effect on stability. In one set of experiments, a powerful, rapidly increasing axial magnetic field was applied to the surface of a short cylinder of sodium. This then produced a bulge in the lines of force, but the cylinder itself was compressed into a tapering billet, the surface of which was lined with prominent furrows. The latter were taken to represent the effect of field lines and fluid interchanging as ripples grew on the surface. In another experiment a modest current was passed down a central conductor through a pot of liquid sodium. The liquid metal was then forced away from the conductor, and under the combined action of magnetic field and gravity, a deep depression was formed about the central rod. The sodium was then frozen, and the new surface was found to be quite smooth.

It is clear that if a fluid is held up against gravity by a horizontal, uniform magnetic field, then a ripple on the surface parallel to the lines of force may have no effect on the magnetic energy nor on the thermal energy, merely interchanging matter and field. When this happens, however, some matter drops through the gravitational field, decreasing its potential energy, and enabling the ripple to grow. This process is the analogue of the Rayleigh-Taylor instability, which occurs when a heavy liquid is supported by a light one, and is the mechanism whereby water falls out of an inverted glass.

This sort of instability has also been demonstrated experimentally, for although the Earth's gravitational field is not strong enough to give a clear demonstration, the acceleration of a fluid surface has the same effect as an oppositely directed gravitational field. If a plasma were accelerated by a rapidly increasing magnetic field, it might be unstable. An experiment performed by a group at Imperial College, London, has shown that this is so and has, moreover, demonstrated that the growth rate has about the predicted value. In this experiment, a cylinder of plasma was suddenly compressed by a rising azimuthal field and was subsequently oscillated in that field. An extremely fast camera was used to make photographs of the plasma surface, which was found to be covered with growing ripples. Measurements of the rate of development of these ripples revealed that in fact they grew only when the plasma was accelerated by the field (figure 9).

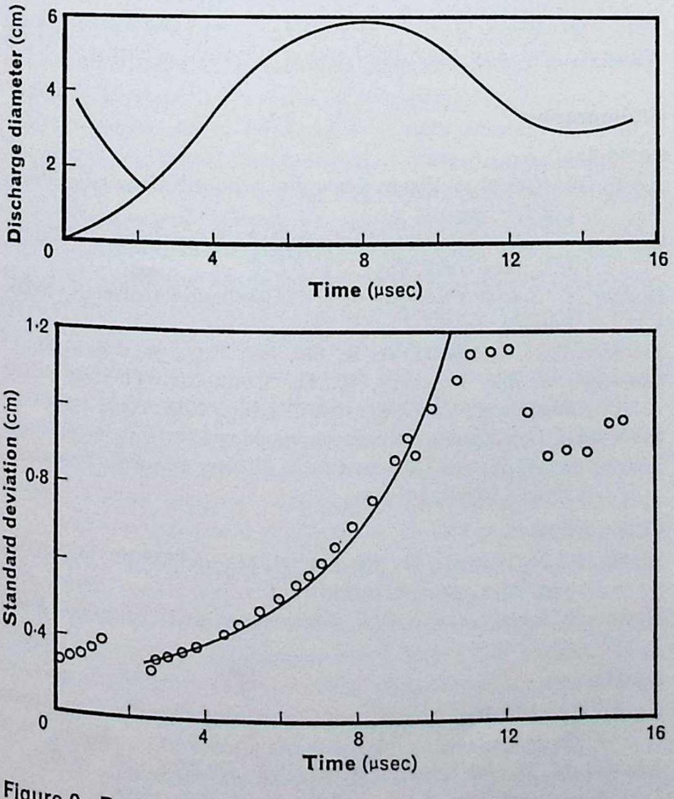


Figure 9 Rayleigh-Taylor instability in a plasma. The upper curve shows the mean radius of a cylindrical plasma as a function of time: when the curvature is negative, the plasma is accelerated by the field. The lower curve shows standard deviation of the plasma radius, made irregular by growing instabilities. Note that this increases only when the field accelerates the plasma (Curzon *et al.*, *Proc. roy. Soc. A*, **257**, 386, 1960).

If a fluid is held against gravity by a field which curves away from the surface, it may be stable, because the loss in gravitational energy when an element of fluid drops is less than the increase in energy of the rising and lengthening field lines. This may explain the relative permanence of the arched prominences that sometimes appear on the solar surface above sunspots.

If the magnetic field is strong and the gas pressure very low, the field will be close to a vacuum field, which is a state of minimum energy, and the energy in the gas may be insufficient to alter that field. In that case, a reduction in energy can occur only if the gas can release thermal energy by some motion in which the magnetic field is unaltered. Conservation of magnetic flux then specifies the change in volume experienced by an element of the fluid when it is displaced. This volume must be that occupied by the magnetic flux originally threading the element of fluid. Just as in the case of thermal convection of the atmosphere, where the stability is decided by the specific entropy gradient and the temperature gradient, so here the system is unstable if the entropy of the fluid decreases in the direction of increasing flux-tube volume. This implies that the plasma is unstable if the pressure decreases rapidly enough in the direction of decreasing field strength, or increasing length of flux lines. There is some evidence from measurements of energy in the Van Allen radiation belts about the Earth, that these are configurations in which the relevant entropy is uniform.

A sad consequence of the stability theory is that the simple confining configurations are, in fact, unstable, because in most of them the field lines are concave towards the plasma surface and field strength decreases outwards. On the other hand, it should be fairly easy to keep a conducting fluid away from the outside of a rigid system of currents. This, while important in some fields, is of little help in research into controlled thermonuclear reactions.

As might be expected, much effort has been expended in pursuit of stable confining systems of a more complex nature: one early approach was that of 'basket weaving'. If the magnetic field near the surface of the plasma suddenly changes direction, then any flux-conserving ripple on the surface must necessarily increase the length of some of the flux tubes, and hence increase the magnetic energy so that the system is stabilized. This method was employed in the 'stabilized' pinch, in which an axial magnetic field is trapped within the plasma. It unfortunately proved unsuccessful because ionized gases do, in fact, have finite electrical resistivity. In order to be completely stable, the configuration must be stable against ripples of any size and shape. For sufficiently small ripples, the scale length may be such that the magnetic Reynolds number vL/η is small even for very small resistivity, η, and very large velocities v. In this case, the resistive dissipation of magnetic

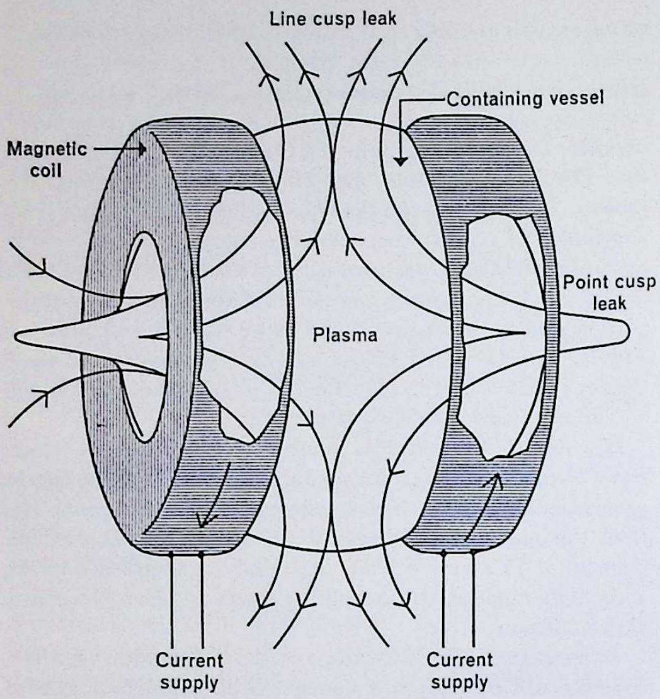


Figure 10 Principle of the cusp.

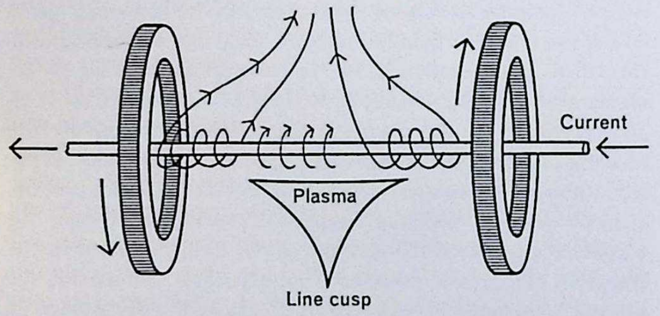


Figure 11 A 'tortured geometry'. The 'stuffed cusp'.

energy cannot be neglected, magnetic flux need not be conserved, and field lines may break and rejoin in the fluid. In fact, the sheet of high current density responsible for the rapid change of field direction can itself break up into a set of little pinches which may move about on the initially smooth surface.

Recently, attention has been turned towards geometries in which the field strength increases as one moves away from the surface of the plasma. A simple geometry which satisfies this condition is the cusp, in which a cylinder of plasma is pinched off by a pair of coils at either end, as in the mirror or thetatron; in this case, however, the currents in the two coils produce opposing fields, which turn away on the midplane, leaving a ring cusp in the field (figure 10). Although stable, this has some bad leaks both at the ends and around the midplane, and is a poor confining system. Experiments have been performed with this geometry at both Los Alamos and Harwell, and those at Harwell have given results, for the containment of a cool plasma, consistent with a stable confinement, although with a diffusive leak through the cusps. Efforts are being made to reduce this loss by increasing the plasma temperature; this reduces the resistivity, and hence the diffusion of field into plasma.

Attempts to improve the cusp geometry have been directed toward adapting it to lower-pressure gases, which may be produced at much higher temperatures, and to closing the cusp leaks. A simple way of doing this is to put a conductor carrying a large current down the axis of the cusp, which helps to close the ring cusp gap: this is the so-called 'stuffed cusp' (figure 11). A second geometry with similar advantages is that of M. S. Ioffe; this comprises a simple mirror to which is added a number of external conducting rods carrying currents in alternate directions, thus producing a configuration in which the magnetic field strength increases outward from the centre. Preliminary experiments by Ioffe in Moscow have produced extremely diffuse plasmas which have been stably confined for long periods, an extremely encouraging result. Many experimenters in all parts of the world are following this lead with great energy and enthusiasm, but while many of these highly tortured geometries appear to be stable, they are extremely difficult to construct, to understand, and to fill with plasma. Hence progress is necessarily slow.

Conclusion

This brief review has mentioned only a few of the ways in which the study of MHD has increased our understanding of natural phenomena, particularly on the cosmic scale, and has emphasized the technological promises. The most attractive of these, controlled thermonuclear fusion, now appears after a number of ups and downs as though it may eventually come true. Meanwhile, research in this field is growing ever more sophisticated, and in some branches is outgrowing the simple language of MHD. Problems are being defined and processes revealed that call for a much more detailed and intimate study of the plasma. Although these more elaborate representations are being developed, MHD retains the great advantage of relative simplicity. Though crude, it is a sufficiently adequate tool for discussing many developments in advanced technology, and for describing such a range of natural phenomena, that some knowledge of it is required by any forward-looking engineer, and any outward-looking physical scientist.

Bibliography

General

Alfvén, H. 'Cosmical Electrodynamics'. Oxford University Press, London. 1950.

Cowling, T. G. 'Magnetohydrodynamics'. Interscience Publishers Inc., New York. 1957.

Dungey, J. 'Cosmic Electrodynamics'. Cambridge University Press. London. 1958.

Frenkiel, F. N. and Sears, W. R. *Rev. mod. Phys.*, **32**, 695, 1960.

Glasstone, S. and Loveberg, R. H. 'Controlled Thermonuclear Reactions', Van Nostrand. Co. Inc., New York. 1960.

Syrovatskii, C. J. Успехи Физических Наук, **62**, 247, 1957.

Thompson, W. B. 'An Introduction to Plasma Physics'. Pergamon Press, Oxford. 1962.

Channel flow

Alpher, R. A., Hurwitz, H., Johnson, R. H., and White, D. R. *Rev. mod. Phys.*, **32**, 758, 1960.

Hartmann, J. and Lazarus, F. *K. danske vidensk. Selsk. Mat-fys. Medd.*, **15**, 6, 1957.

Equilibrium

Grad, H. and Rubin, A. *Proc. 2nd Geneva Conf. Peaceful Uses Atomic Energy*, **31**, 190, 1951.

Kruskal, M. D. and Kulsrud, R. M. *Phys. Fluids*, **1**, 265, 1958.

Lüst, R. and Schlüter, A. *Z. Naturforsch.*, **12**, 850, 1957.

Laing, E. W., Roberts, S. J., and Whipple, R. T. P. *J. nucl. Energy*, **1**, 49, 1959.

Stability

Bernstein, I. B., Frieman, E. A., Kruskal, M. D., and Kulsrud, R. M. *Proc. roy. Soc.*, *A*, **244**, 17, 1958.

Chandrasekhar, S. 'Hydrodynamic and Hydromagnetic Stability'. Oxford University Press, London. 1962.

The movement of soil water

E. C. Childs

A study of the movement of water through soil, considered as analogous to the flow of heat through a conductor, was developed more than a hundred years ago. This treatment applies only to the simplest cases, because the hydraulic conductivity of a particular specimen of soil depends in part on its history. This article describes the extension of the simple theory to more complex cases, in the laboratory, and to the practical study in the field of land drainage and the irrigation of plants.

In a brief discussion of a broad topic, we must simply accept that soil is one particular example of a class of porous materials in which the pore space is continuous. The soil properties that we shall examine all spring from the geometry of the pore space, which may depend systematically upon the depth from the surface of the land. The variation of a property with depth is usually referred to as a profile.

Among the many constituents of soil are two that are major contributors to its characteristic properties. These are the clay fraction, which consists largely of a few classes of layer lattice minerals known collectively as the clay minerals, and the soil organic matter or humus. Biological activity and the weathering of the parent rocks, both of which are less intense at greater depths, together with the carriage of the products by downward percolating water, cause the development of profiles of the clay fraction and of humus. These two kinds of material have chemical structures which, for different reasons, may release ions, usually cations, into a solution with which they are in contact so that they themselves remain with negative electric charge. This charge is the origin of the colloidal properties which soils exhibit and which are fairly well understood. For an exposition of this subject one may refer to a publication by E. J. W. Verwey and J. Th. G. Overbeek [1].

Notable among these colloidal properties is the tendency of particles to remain in stable dispersion when suspended in dilute solutions of monovalent ions such as sodium, and to flocculate in more concentrated solutions, particularly of polyvalent ions. A second property, which is also sensitive to the concentration and valency of ions in the solution in contact with the surfaces of the particles, is the power of masses of the material to imbibe water forcibly with consequent swelling in spite of the application of confining pressures. There is a third property associated with the colloidal fractions, and particularly with the biological activity which results in the formation of humus. This is the association of the ultimate soil particles in aggregates that have some claim to persistent individual existence in the face of disruptive forces. The shapes, sizes, and mutual arrangements of aggregates define an inter-aggregate pore space which has a geometry on a larger scale than that of the inter-particle pore space within the aggregates. Even though the coarse structural channels may contribute little to the total volume of pore space, their contribution to water movement may be dominant. Since structural fissuring may develop to different degrees in different directions, the ability of soil to conduct water may also differ in different directions. In such cases, the conductivity is said to be anisotropic.

The development of a more or less stable soil structure is not well understood and is certainly not synonymous with flocculation of soil colloids. Nevertheless, it would be surprising if stable structure formation were not inhibited by circumstances favouring dispersion, and in fact this incompatibility of structure and dispersion is found in nature.

E. C. Childs, *B.Sc., Ph.D., Sc.D.*
Was born in London in 1907, and educated at East Ham Grammar School, University of London (King's College), and University of Cambridge (Cavendish Laboratory). He is now Reader in Soil Physics in the Department of Agriculture, University of Cambridge, and Honorary Director of the Agricultural Research Council's Unit of Soil Physics.

Soil water

The density and viscosity of a liquid together determine how fast it will flow through a given channel with a given pressure difference between its ends. Neither density nor viscosity is very profoundly affected by dissolved salts in low concentrations. The so-called soil water is in fact a dilute solution of a great number of different salts, but its density and viscosity can with little error be regarded as those of pure water.

Although the constitution of the soil solution does not affect its density and viscosity, it does affect the stability of the soil structure. Commonly, the divalent calcium ions predominate, so that even in dilute solution there is only a small tendency towards dispersion. However, mishaps such as a catastrophic sea invasion or ill-advised irrigation may cause a predominance of monovalent sodium ions in high concentraion. Later, when the concentration is reduced by percolating rainwater without reduction of the relative predominance, the result is often a dispersion of the clay and destruction of the soil structure. Whatever may be said later about soil properties must not be held to imply that these are immutable, for they may evidently be changed according to the composition of the percolating solution.

The static equilibrium of soil water

There are two mechanisms by which soil water is retained in spite of attempts to extract it by suction. If the soil does not shrink, then the removal of water must be accompanied by the entry of air, with the formation of a curved interface between air and water in the pore space. Surface tension acting in this interface supplies the pressure difference needed to balance the extracting suction. The smaller pores impose sharper curvatures of the interface than do the larger, with consequently greater pressure differences, so that with increasing suctions the interface retreats into smaller pores with increasing loss of water. The curve of moisture content c plotted against the negative (in relation to atmospheric) hydrostatic pressure p applied to the soil solution is called the moisture characteristic.

In colloidal materials, the suction is balanced by the mutual repulsion of the surface-active particles. With increasing suction, the particles are pulled closer together, producing loss of water with an equivalent shrinkage but without air entry. The moisture characteristic naturally differs in shape from that which relates to air entry. Structured soils commonly have moisture characteristics which at low suctions reflect air entry into the structural pore space, but at higher suctions correspond to shrinkage of the aggregates.

We need to note two particular features of the moisture characteristics. Firstly, they are not straight lines, and secondly, they exhibit hysteresis. The drying characteristic does not retrace the path of the wetting characteristic, and the two together form a closed loop. Characteristics that are due to reversals of trend between the limits are curves that lie within the primary loop. Thus the suction and the rate of change of moisture content with suction are not uniquely prescribed when the moisture content or suction are specified, for the history of wetting and drying is relevant. We note also that the curves are not straight lines, so that the slope dc/dp is not a constant, characteristic of the material.

A. Poulovassilis [2] pointed out that hysteresis has a further consequence. The configuration of the water-filled pore space is not uniquely prescribed by the moisture content, so that the paths available for the flow of soil water are also affected by the past history of the sample. The effect is, however, too small to be demonstrated with confidence.

The conduction of water in porous materials

By analogy with the conduction of electricity, we may define hydraulic potential in the following terms. The work done against the prevailing field forces in transferring unit quantity of water from point *A* to point *B* is the amount by which the potential φ_B at *B* exceeds that, φ_A, at *A*. The unit quantity may be chosen for convenience, but if unit weight is chosen, then it turns out that φ is the sum of two terms thus:

$$\varphi = z + p$$

where z is the height, above an arbitrary datum, of the point where φ prevails and is the gravitational term, and p is the height of water in the limb of a manometer inserted at the point and represents the term due to hydrostatic pressure. If p is negative, that is to say a suction, then the manometer must be of a special kind called a tensiometer.

In 1856, H. Darcy [3] published *Les fontaines publiques de la ville de Dijon*. This contained a discussion of the vertical filtration of water through sand beds and enunciated a law analogous to that presented some thirty years earlier for the flow of heat in a thermal conductor. The law may be generalized and expressed in terms of potential in the form known as Darcy's law,

$$v = -K \operatorname{grad} \varphi$$

The gradient of potential is the increase of potential per unit distance in the direction in which the potential increases most steeply. The velocity of flow v is determined from the measured rate of flow by supposing that the flux of water is quite uniformly distributed over the whole area of cross section. Darcy's law would not, therefore, necessarily hold if the sample were so small as to contain only a few unrepresentative channels.

The parameter K is the hydraulic conductivity of the material. If this is anisotropic, it has the effect of giving v a direction which differs from that of the potential gradient, and K will be a tensor. However, this difference need not be stressed, because simple mathematical devices are available whereby a problem with anisotropic conductivity may be converted to an equivalent isotropic case in which K may be treated as a simple scalar factor.

Darcy's law is formally identical with the law of thermal conduction which is the foundation of an extensive body of analysis of heat movement. We might expect this analysis to be applicable at once to water movement, but unfortunately both K and φ present features which have no thermal analogues. Firstly the potential has a gravitational term, namely the height in space, as well as one representing the hydrostatic pressure. It is as though the temperature of a body were to depend upon its height as well as upon the reading of the thermometer. Secondly, because water can flow only through water-filled channels, which become not only fewer but also smaller as the moisture content decreases, the conductivity (K) is greatly dependent upon moisture content, c. A typical variation of K with c is shown in figure 1. The moisture content is related to suction which in turn is a contribution to the potential, so that K is a function of φ in a given situation.

Because even a small proportion of the pore space devoted to a few large channels contributes far more effectively to the conductivity than does a large amount of finely divided pore space, the conductivity of soil is greatly dependent upon the development of structure, and this is particularly true of heavy clay soils, the conductivity of which would often be negligibly small in the absence of structural features. The conductivity may thus be reduced markedly by any mismanagement that results in the destruction of structure, and such mismanagement includes salinization.

The measurement of the hydraulic conductivity of both saturated and unsaturated samples using laboratory techniques following thermal analogy is quite possible for structureless bodies, but not for soils. This is because the removal of a sample so modifies the all-important structure that the results bear no relationship to the conductivity in the undisturbed field state. There are satisfactory methods available for the measurement of the conductivity of saturated field soils where groundwater occurs, based on the observed relationship between the rate of pumping of water from wells, boreholes, and cavities of various kinds, and the level of water in them.

The solution of problems of heat flow is often facilitated by expressing the law of thermal conduction as a proportionality between the rate of flow of heat and the gradient of heat content in the conductor. This is possible because the specific heat relates this gradient to the temperature gradient. As thus stated, the law is formally identical to the law of molecular diffusion, with a constant of proportionality, or diffusivity, given by the ratio of the thermal conductivity to the specific heat. In hydraulic conduction, however, this procedure is not always fruitful. It may be that different points on the moisture profile have had different histories of wetting and drying, and in that case a component of the moisture gradient may be due simply to this difference, contributing nothing to the suction gradient. Even that component which provides a related suction gradient, via the slope of the moisture characteristic, presents a diffusivity which is moisture-dependent, since both the conductivity and the slope of the moisture characteristic, of which the diffusivity is the ratio, are themselves dependent upon the moisture content.

In all but the simplest situations, progress is made by appeal to the equation of continuity, which is a mathematical formulation of the statement that the net excess of the rate of flow of liquid out of an element of the body equals the rate of loss of moisture by that element. The rates of flow everywhere are expressible in terms of the potential gradients according to Darcy's law, so that the equation of continuity relates the change of moisture content everywhere to the distribution of potential. In vector terminology it takes the form

$$\partial c / \partial t = \operatorname{div} (K \operatorname{grad} \varphi)$$

Of this is sufficient to say here that it is a differential equation which describes flow generally, and that particular situations are described in terms of known conditions of certain of the variables at known boundaries. If a solution is then possible, it describes the distribution of the variables throughout the rest of the body. The procedure is best illustrated by considering some particular classes of situation.

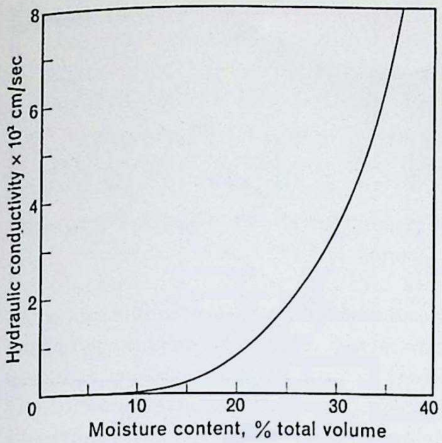


Figure 1 Example of variation of hydraulic conductivity with moisture content

The soil moisture profile

A general class of problems deals with the distribution and movement of water vertically down the soil profile and with changes in the moisture profile with the passage of time. With only the z-direction to be considered, the equation of continuity takes the form

$$\partial c/\partial t = \partial[K(1+\partial p/\partial z)]/\partial z$$

An important sub-group of such problems includes all those where moisture moves at a steady rate up or down the profile from or to a water table without engendering a change of shape of the moisture profile. Elimination of time as an independent variable leaves the equation of continuity as an ordinary differential equation of two variables which may be integrated, in the last resort by numerical or graphical methods, provided that K is known as a function of moisture content and of suction. The result is in the form of a distribution of moisture content or of suction above the water table, or, in other words, a moisture profile.

In this way one can discover how the moisture profile adjusts itself to accommodate increasing rainfall rates, so as to store water that is released when the rain stops, This calculation can be made even when the water table is falling or rising steadily, provided that the moisture profile has time to settle down to a steady shape which rises or falls with the water table. Again, it may be shown that steady upward movement of water from a water table to make good evaporation from the land surface can take place only from a water table within a limited distance from the surface, and this distance may be ascertained and shown to depend upon the rate of evaporation.

In another class of situation, we require to follow the development of the moisture profile with passage of time as infiltration takes place into an initially dry soil from a flooded surface. In particular, the rate of increase of the total amount of soil water so revealed is a measure of the rate of infiltration. This is of direct importance when irrigation has to be controlled to a timetable specified by the water supply authority, as is commonly the case in the major irrigation areas, and is of indirect but no less importance in areas which are susceptible to soil erosion, for it is the arrival of excess water at the surface beyond the infiltration rate which gives rise to surface run-off and consequent erosion.

This type of problem is best studied by expressing Darcy's law in terms of diffusivity, as outlined earlier. The requirement that the history of the moisture profile shall everywhere be the same is effectively met if the initial state is one of such dryness that the effects of previous cycles of moisture change have been erased. The equation of continuity then takes the form

$$\partial c/\partial t = \partial[K + D(\partial c/\partial z)]/\partial z$$

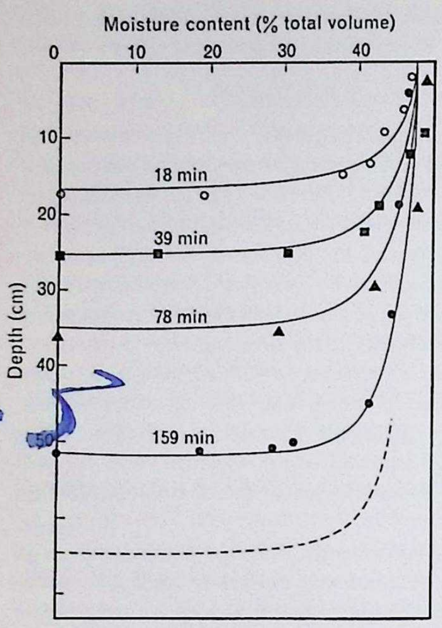
where D is the moisture-dependent diffusivity. The gravitational term K has no thermal analogue and there are for that reason no ready-made solutions, but J. R. Philip [4] has provided one in the form of a series of terms each of which has a coefficient which is itself the solution of one of a set of auxiliary approximate equations. It proves to be suited to describe only the earlier stages of infiltration, but a transformation first proposed by S. Irmay [5] again eliminates the time variable to provide a solution in the form of a moisture profile of calculable shape advancing at a calculable speed, and this describes all late stages. E. G. Youngs [6] made calculations of these kinds based upon a specified material which was also the subject of experimental observations, with confirmatory results which are shown in figure 2.

These analyses and experiments, difficult as they are, are as yet applicable only to uniform materials and laboratory studies. They serve to elucidate the physics of the subject but do not provide an effective means of field investigation. Fortunately, it is found that infiltration rates computed by adopting an older and much more naïve model introduced fifty years ago by H. Green and G. A. Ampt [7] are not seriously at variance with those predicted by the more rigorous modern analysis in those cases where a comparison may be made. These authors considered the development of the moisture profile to consist in the advance of a water front that separates completely saturated soil from that as yet unaffected by the flooding. At this precisely located front, a characteristic suction is supposed to act. Thus at all stages, even in stratified soil, we have a known potential difference applied between the ends of a column of soil of known resistance to flow, so that the flow rate, which equals the infiltration rate, may be computed by a direct application of Darcy's law.

Land drainage

Water may enter a land area by a number of routes, including local rainfall, the flow of surface or ground water from a foreign source, and the entry of artesian water. Similarly, it may leave by natural seepage, by evaporation at the surface or through vegetation, and by flow through natural or artificial drainage channels. All this involves the establishment of a potential distribution in conformity with Darcy's law and this again implies the establishment of a pressure distribution. In general, deeper zones in the soil profile experience a positive hydrostatic pressure, and suctions prevail nearer the surface. The surface at zero suction separating these two zones is the water table, and the soil profile is commonly saturated for a certain distance above the water table, within which distance the suction is insufficient to initiate desaturation. This zone of persistent saturation under suction is called the capillary fringe.

If for any reason the water table is unacceptably high, then one must resort to artificial water exits of one sort or another, such as open ditches, underground drainlines, or pumped wells. These act by imposing near-zero pressures at these selected surfaces, and this so modifies the pressure distribution everywhere as to bring the moisture content to within acceptable limits. An example is shown in figure 3, where a set of parallel equidistant drains at uniform depth is seen controlling a water table to an acceptable height H in spite of rainfall at a rate q. The problem here is to discover the distribution of potential when the relevant parameters, including the soil



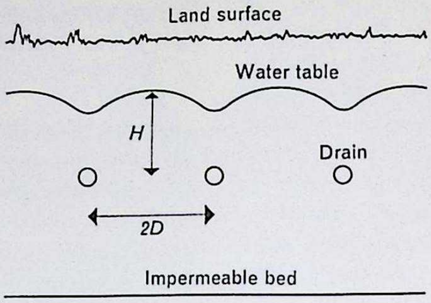


Figure 3 A section of a drainage system of parallel equidistant drains controlling a water table due to local rainfall.

Figure 2 (left) Moisture profile stages during infiltration from a flooded surface. Solid lines are theoretical curves, after J. R. Philip. The marked points are experimental observations. The broken line is the theoretical curve for late stages, after S. Irmay.

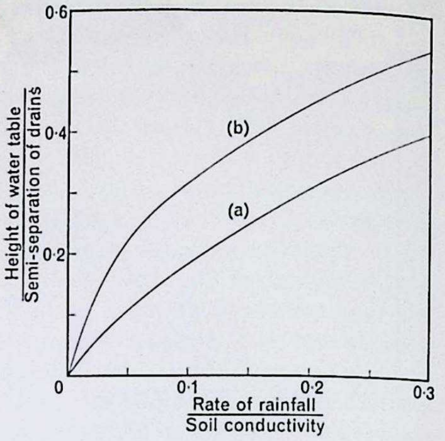


Figure 4 Theoretical curves of the relationship between *H*/*D* of figure 3 and the ratio of rate of rainfall to hydraulic conductivity of the soil. Curve (*a*) is for a deep impermeable substratum, while curve (*b*) is for an impermeable bed at drain level.

conductivity K and the drain separation $2D$, are known. In particular, the location of the water table will emerge as the locus of points where the hydrostatic pressure is zero.

It must suffice here to state that, in general, methods of solution of the equation of continuity for drainage situations are limited to the steady state, although some methods may be adapted to treat a rising or falling water table as a sequence of stages each of which is momentarily a steady state. A thorough treatment of the subject is available as a monograph edited by J. N. Luthin [8].

When any impermeable substratum which may be present is at great depth, there is available an analytical solution from which the curve (*a*) shown in figure 4 has been drawn. When the impermeable bed is at drain level one may employ the Dupuit-Forchheimer convention, which supposes that the flow is everywhere horizontal and therefore takes place between vertical equipotential planes. This proves to be acceptable in more situations than one has any right to expect, and in the present case leads to a simple formula which has been used to draw the curve (*b*) of figure 4. Intermediate cases may be treated by interpolation, but any particular case may be examined by guessing a potential distribution and correcting it by a systematic reiterative procedure until it fits the equation of continuity and the boundary conditions. This procedure is tedious, but the work may be programmed for performance by an analogue or digital computer.

Soil water and the plant

An upper soil zone containing plant root systems may be regarded as drawing water from a lower zone which lacks these roots in much the same way as the atmosphere draws water from the upper soil. In just the same way also the depth which may be drawn upon is limited. The recognition of this phenomenon is expressed in the commonly heard sentence that roots must grow to the water, for the water will not come to the roots.

This must not be interpreted as implying that the root must invade every pore in order to exploit the contained water, for within the root zone, the extraction of water by the root can set up very steep potential gradients over the short distances that are operative. On the other hand we must exercise caution in applying Darcy's law to this situation, because the sample in the neighbourhood of the root is generally too small to apply this statistically based relationship. One can do no more than argue qualitatively, but even at this level much light is thrown on such phenomena as the wilting of plants and the germination of seeds in soils at various moisture contents.

By absorbing water in their immediate vicinity, such organisms desaturate the soil locally. This introduces potential gradients on a micro scale, which tend to maintain water movement toward the organisms, and at the same time reduce the hydraulic conductivity locally, thus hindering water movement. It is this combination of effects which is important, and not merely the soil suction experienced by the surface of the root or seed, and certainly not the average soil suction as measured in the sample. Thus J. R. Gingrich and M. B. Russell [9] compared the rate of elongation of the maize radicle in a soil under hydrostatic suction with that in a solution with the equivalent osmotic pressure, and demonstrated marked differences. This wide field of the physics of the interaction of soil water and organisms is still, because of the great difficulties, very little explored.

References

[1] Verwey E. J. W. and Overbeek, J. Th. G. 'Theory of the Stability of Lyophobic Colloids'. Elsevier Publishing Co. Inc., New York. 1948.
[2] Poulovassilis, A. *Soil Sci.*, 93, 405, 1962.
[3] Darcy, H. 'Les fontaines publiques de la ville de Dijon'. Dalmont, Paris. 1856.
[4] Philip, J. R. *Aust. J. Phys.*, 10, 29, 1957.
[5] Irmay, S. *Proc. Symp. Darcy Int. Assoc. Sci. Hydrol. Publ. Ser.*, 41, 57, 1956.
[6] Youngs, E. G. *Soil Sci.*, 84, 283, 1957.
[7] Green, H. and Ampt, G. A. *J. agric. Sci.*, 5, 1 1912.
[8] Luthin, J. N. (Editor). 'Drainage of Agricultural Lands'. American Society of Agronomy, Madison, Wisconsin. 1957.
[9] Gingrich, J. R. and Russell, M. B. *Soil Sci.*, 84, 185, 1957.

Germination of fungus spores

David Gottlieb

Our concept of the mechanism of spore germination has recently changed. Absorption of water, with consequent swelling, causing the cell-wall to rupture, does not wholly explain the process. Before any germ tube is produced, a very complex series of biochemical reactions occurs, involving the production of energy from carbohydrates, synthesis of enzymes, and many other changes.

Fungi are familiar as the edible mushroom, the bracket fungi that grow on trees, the moulds that spoil our food, or the causal agents of certain plant or animal diseases. In all, there are some 40 000 species. No matter what their morphological appearance, the organisms are essentially unicellular, and each cell is normally metabolically independent. If separated from the other cells in the body of the fungus, almost any single cell can develop all stages in the life cycle of the fungus. In the majority of fungi it grows into a thread-like hypha that branches and eventually forms a mat of hyphae, the mycelium. In some species, notably the yeasts, the cells are ovoid and grow by budding off new cells that may separate or remain attached.

The nutritional requirements of most fungi are relatively simple, though some, such as the obligatory parasites, are exceptions. Their elemental needs can be satisfied with minerals, oxygen, and an organic source of carbon. The requirement for organic carbon differentiates them from green plants, which use the carbon dioxide of the air to photosynthesize the cellular constituents. Some species of fungi lack the ability to make all the metabolites needed for their activities, such as biotin and thiamine, and can grow only when preformed sources are available.

The process of reproduction differs among the many species, but has some common features. At certain stages in the growth of hyphae there is a morphological change, and specialized hyphae that function in reproduction appear. One of two types of spore can be produced: sexual, which require a mating process, or asexual, which do not require an exchange of genic material. Sexual spores make genetic variation possible in the species; they are usually heavy-walled to allow the new organism to survive rigorous environments. Asexual spores arise by a segmentation of specialized hyphae, which may be simple erect hyphae or grouped to form special reproductive bodies. Spores may also be produced within a container, the sporangium, inside which the protoplasm is divided up into zoospores and aplanospores. Motile spores are adapted to moist environments and resemble certain protozoa; they have flagella and propel themselves through the liquid before settling down to produce the adult plant. Spores can be uninucleate or multinucleate; single-celled or many-celled; thin-walled or thick-walled. Spores known as chlamydospores are especially adapted for survival in adverse conditions, and are simply hyphal cells that have acquired a very tough cell coat.

All types of spore have some adaptation for release and dispersal. Some species, such as *Penicillium* species, have thick-walled spores that are easily carried by the wind and resist desiccation. Others, such as *Phytophthora infestans*, have thin-walled spores that are also carried in the air but require humid conditions for survival. Spores of *Aphanomyces cochliodes*, a soil-borne parasite, swim through water to find new hosts. Each member of the species must finally lodge in an environment that is not occupied by other species, or it must be able to overcome an established competitor. Sometimes, however, a close association of two or more organisms favours survival when one member can supply the nutritional needs of another. After the spores have been produced, and during their period of dispersal, the spores are dormant; no growth occurs, and their metabolic rate is extremely low. Many more of these microscopic spores are produced than ever reach a suitable environment, and relatively few succeed in germinating.

Morphologically, the dramatic transformation from dormant to active mycelium can readily be seen with the microscope. In a suitably moist environment, the spore swells, sometimes to eight times its original volume. Later, a slight bulge occurs in the spore wall. The spore wall is ruptured, and the cell that had been confined expands into a peg-like protuberance which further extends and forms the filamentous germ tube (figure 1). In due course the germ tube becomes a vegetative hypha, the beginning of the obvious body we call the fungus. Germination has been defined as the appearance of the germ tube.

Obstacles to germination

Sometimes spores do not germinate even under good nutritive conditions. Laboratory experiments with large numbers of spores led to the observation that in rust and powdery mildew species the percentage of spores that germinated under crowded conditions was very low. Diluting the spore suspension resulted in an increasing percentage of germination [1]. In the rust fungus *Puccinia graminis tritici*, this type of inhibition is caused by the production of toxic material by the spores; a high spore density results in high toxin concentrations. The inhibitor can be leached out with water. It is volatile, and its accumulation can be prevented by aerating the solution; most evidence indicates that the toxin is trimethylethylene [2]. Other metabolites that have been implicated are glutamic and glutaric acids. On the other hand, no germination may occur at very low spore concentration, while at slightly higher levels the percentage germinating rises. Some still undefined substances are evidently needed if germination is to proceed. Both types of agents, stimulating and inhibiting, can be produced at the same time [3]. The natural significance of these laboratory phenomena remains to be determined, for it is doubtful whether the number of spores landing on any substrate is great enough to allow either inhibitory or stimulatory materials to become concentrated. The very fact that spores are carried to a new substrate by air currents—in the case of the rusts, to a new plant—also implies enough air movement to disperse volatile compounds that might otherwise accumulate.

Sometimes spores are not physiologically mature when they are deposited on a substrate and so cannot germinate. Many examples of such phenomena have been reported, such as the telospores of the rust fungi, ascospores of *Neurospora tetrasperma*,

David Gottlieb, *B.S., M.S., Ph.D.*
Was born in New York City in 1911 and is a graduate of City College, Iowa State College, and the University of Minnesota. He is now Professor of Plant Pathology at the University of Illinois. His researches have been in three main fields. Firstly, antibiotics — their mode of action, biosynthesis, and search for new antimicrobial agents; secondly, the role of antibiotics in soil; and thirdly, the biochemistry of fungi.

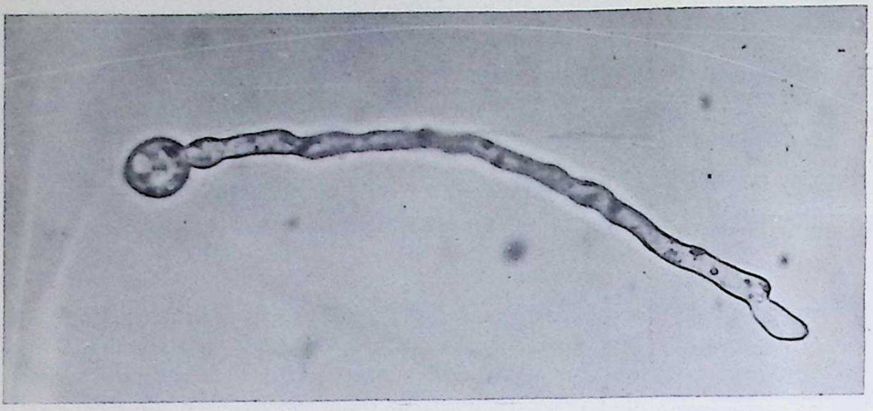


Figure 1 A spore of *Aspergillus flavus* incubated in a nutrient medium, showing formation of the germ tube.

and the sporangiospores of *Phycomyces blakesleeanus*. With long lapse of time and variation of the environmental conditions maturation eventually occurs. Telospores of *P. graminis tritici* collected from wheat plants in the late summer do not readily germinate, but if left on the plants in the field until the next spring they will do so. Simulation of the conditions occurring in nature—such as alternate periods of wet and dry, hot and cold—stimulate germination under laboratory conditions. Short periods of heat treatment—about 50°C for a few minutes—break the latent period of spores of *Phycomyces* and *Neurospora*. Treatment with various chemicals, such as furfuraldehyde, pelargonaldehyde, and many other organic compounds, promotes germination of rusts. Formerly, these effects were attributed to changes in the permeability of the spore to water, but lately the stimulatory activities of these treatments have been correlated with their ability to increase the oxidative activities of the cell [4]. D. R. Goddard's early studies of heat activation of *Neurospora* ascospores had similar implications [5]. The glycolytic system was first stimulated; later, on formation of the germ tube, an aerobic type of oxidation became prominent.

Water absorption hypothesis

For many years, germination was looked on as a relatively simple process. The spore cells were regarded as containing a desiccated colloidal protein complex surrounded by a cell membrane and protective wall. The moisture content is always much lower than that of mycelium of the same species and usually varies from 6 to 25 per cent; occasionally, as in the powdery mildew, higher values are found. In the absence of free water or high humidity, physiological reactions were scarcely perceptible. The colloid, it was supposed, would absorb water and swell to many times its original volume, until finally the pressure on the spore wall became so great that it burst and the cell contents escaped. This escaping cell material formed the germ tube. Though the dependence of the process on many factors—such as age of spores, high humidity, temperature, and hydrogen-ion concentration—were well documented [6], these effects could be related to colloidal phenomena or simply left unexplained. Nevertheless, the results of such environmental effects were recognized as being of very great practical importance, for they were relevant to the control of fungal infections of both plants and such valuable property as food and clothing.

When old data on spore germination, on which such beliefs were founded, are examined in the light of current knowledge, the need for explaining germ-tube formation in metabolic terms is obvious. A slight oxygen consumption was recorded as occurring even in dry spores. It increased when the spores were immersed in water, was further stimulated by the presence of nutrients, and then sometimes increased again when the germ tube appeared [7].

Another indication of metabolic activity apparent in the early records was the change in pH which occurred when spores were placed in various media; such changes are similar to those of growing cells. Careful studies had shown the need of certain specific nutrients for germination in some species. Among these were inorganic nitrogen and magnesium compounds; organic compounds, especially simple sugars; and sometimes vitamins and amino acids. Sometimes, unidentified natural products of a fairly complex character also seemed necessary. In one of our own recent studies with *Penicillium oxalicum*, the following chemicals were found necessary for optimum germination: glucose, asparagine, phosphate, and magnesium sulphate. Ammonium nitrogen could partly take the place of asparagine nitrogen, but nitrate was ineffective. Oxygen and water were also needed, so that the macroelements required for growth of the mycelium were also required for spore germination.

Composition of spores

The gross chemical composition of spores resembles that of hyphae of the same genus. Carbohydrates constitute from 20 to 35 per cent dry weight, lipids between 12 and 35 per cent, and protein about 25–30 per cent. The actual composition will vary with the nutrient medium on which the fungus was grown. A recent analysis of spores of *Neurospora* [8] gave the following results, expressed as per cent dry weight: ash 6·8, protein 25·8, lipid 17·7, ribose nucleic acid 9, deoxyribose nucleic acid 1·4, carbohydrates 9·3, organic acids 0·2, free amino acids and amides 1·3. The carbohydrates and related substances included mannose, glucose, and arabitol. Trehalose has also been found in spores, and may be one of the main storage forms of glucose. Chitin is probably present in the spores of most fungi. It has been demonstrated in cell walls of species of the four main groups of fungi, the Phycomycetes, Ascomycetes, Basidiomycetes, and Fungi Imperfecti. The fungi in which it has not been found are the Oomycetes, a division of Phycomycetes, and basidiospores of *Ustilago zeae*. The cell walls of the Oomycetes contain cellulose in place of chitin. In spores of *Neurospora tetrasperma*, 8 per cent of the total cell-carbon is in the metabolic pool and thus ready for immediate use for either energy derivation or synthesis of cell components.

The most important functional units in the spore are the enzymes that catalyse the various metabolic reactions. As long ago as 1923, a paper was published entitled 'Do mould spores contain enzymes?' [9]. The question was answered in the affirmative, for a few hydrolytic enzymes had been demonstrated. Today, a long list of spore enzymes can be compiled. As will be discussed later, all the enzymes known to be needed for the catabolism of glucose and for respiration have been found. Polyphenol oxidases have been shown to be present in most species. Ascorbic oxidase, sulphydryl oxidase, pentose reductases, and others have been reported for particular species. One would assume the presence in spores of almost all the usual metabolic enzymes, were it not that occasionally we find an organism whose spore can carry out a reaction that the vegetative mycelium cannot, or vice versa. Most of the enzymes are located inside the cell, either as colloids in aqueous suspension or more intimately associated with particulate cell fractions, as, for example, some electron transport enzymes are with the mitochondria. A few, however, seem to be associated with the external surfaces of spores.

Metabolic basis of germination

That metabolic activity is intimately associated with spore germination is indicated by the fact that changes in chemical composition occur during the process. In *P. graminis tritici* and *Uromyces phaseoli*, for example, there is a decrease in lipid content of the spore on germination. In various species of Saprolegniaceae the oil droplets disappear during germination. In other species the amount of carbohydrate decreases. Species that cannot utilize their internal carbon or lipids and depend on external sources for their carbon requirements show no overall decrease. Phosphate and chitin losses have also been shown in some species. Such changes can occur very soon after the spore is placed in the germination medium and probably do not lag much behind increased oxidation. During its swelling stage, *Aspergillus niger* slowly synthesizes nucleic acids, free amino acids, and protein; and oxygen consumption rises at the same time. The rate of protein synthesis and oxygen consumption increase further when the germ tube begins to form [10]. *A. nidulans* accumulates deoxyribose nucleic acid and protein within four hours of incubation; the protein increase is probably at the expense of free amino acids, which decrease. Later, ribose nucleic acid increases rapidly, and at germ-tube formation the rate of synthesis is still higher [11].

A need for organic carbon is generally satisfied simply by glucose; our own studies indicate that such a need is the rule rather than the exception. If there is no apparent need for organic carbon in the medium, one can reasonably assume that there are reserves in the spore that can be used for the purpose. A number of studies have shown that many spores have all the enzymes needed for the catabolism of glucose. Enzymes necessary for an aerobic pathway leading to acetaldehyde have been demonstrated during germination in four fungi that we have studied, *P. graminis tritici*, *Uromyces phaseoli*, *Ustilago maydis*, and *Penicillium oxalicum*. In addition, *P. oxalicum* also was found to contain alcohol dehydrogenase [12]. Also present are all the enzymes of the aerobic pentose pathway for the breakdown of glucose. With two exceptions, the spores contained a full complement of the enzymes needed for the more complete oxidation of the intermediate products to carbon dioxide via the Krebs cycle. Probably, the missing enzymes (pyruvic dehydrogenase and α-ketoglutaric dehydrogenase) were originally present in the spore, but were inactivated. Extracts of the mycelia of these fungi never showed these activities, despite variations in extraction techniques. Other dehydrogenases were present in ten fungi that were examined. The spore had a complete terminal electron-transport system which carries out the final step in the oxidation of sugar, the transport of electrons to oxygen.

The importance of an external carbon source for respiration in *P. oxalicum* has been shown experimentally. Spores that received no glucose consumed very little oxygen and oxidized whatever substrate was available within the spore. Eventually even that was not enough, and respiration dropped to a negligible level. Spores that had been previously incubated in glucose solution, so that the sugar entered the cell, consumed four times as much oxygen as untreated ones. This treatment introduced enough carbohydrate into the spore to keep respiration and synthesis going for at least 24 hours. The rise in oxygen consumption between the 6th and 12th hours corresponded to the period of germ-tube formation. But one must not draw the conclusion that all carbon sources which stimulate respiration also stimulate germination; acetate, for example, stimulates respiration in this organism without effecting germ-tube formation.

The obligate parasites *P. graminis tritici* and *Uromyces phaseoli* present a somewhat different picture of carbon metabolism. Their uredospores need no external carbon for germination, and form germ tubes in distilled water alone. They can progress no further, however, unless the germ tube penetrates the host plant. Since germination is sharply delineated from the growth function, these organisms should present the simplest example of water absorption, swelling, and germ-tube formation. Again, increased oxidation in water suggests complex metabolic activity, and these species have all the many enzymes required for carbohydrate metabolism that have been demonstrated in other spores [12]. However, respiration is not at the expense of either internal or external carbohydrates, but at the expense of the lipids that have been stored in the uredospores. As shown by radioactive tracers, carbohydrate catabolism is very weak and the addition of glucose fails to stimulate oxygen consumption. On the other hand, lipids in the spore decrease during germination (figure 2) and the presence of fatty acids stimulates respiration [13]. The respiratory quotient of rust spores is typical of a fatty respiration. Furthermore, the saponification number of the lipids in the spore rises during germination, indicating the hydrolysis of fats to fatty acids. Acetate stimulates respiration more readily, and at an earlier period, than does glucose. After hydrolysis of the glycerides, acetate is probably derived by a β-oxidation of the fatty acid, and this metabolite can be oxidized in the Krebs cycle with release of energy and formation of intermediates for amino acid synthesis. All these facts are consistent with a fatty acid metabolism being dominant during the germination of rusts.

Synthesis of cell components

While the spore is being incubated, it is also synthesizing cellular material from either its own reserves or extra-cellular

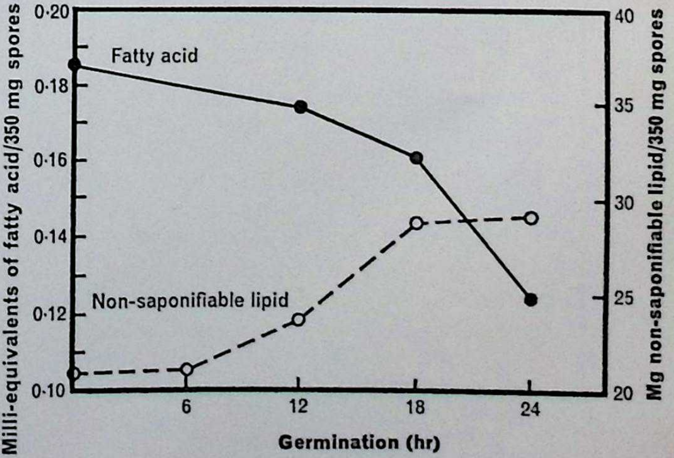


Figure 2 Decrease in fatty acid and increase in non-saponifiable lipid during germination of uredospores of *Puccinia graminis tritici*.

sources. *Penicillium oxalicum* forms cellular carbohydrates from either sugar or acetate, as shown by the incorporation of carbon-14-labelled substrates. Similarly, ^{14}C-acetate is synthesized into carbohydrate by *P. graminis tritici* (figure 3). In both, the total carbohydrate is increased when they germinate in water. This synthesis probably arises from the stored lipids [14]. From acetate, glucose is the sugar most readily formed, followed by mannose, fructose, and arabinose; trehalose has also been reported [14]. Whether or not there is a net synthesis of protein in the rusts is still debatable. Our investigations indicate some overall synthesis during the early stages of germination, but the results of other studies do not. There is no doubt, however, that some protein is being synthesized, for acetate labelled with carbon-14 has been shown to be incorporated into protein.

Very little is known about the enzyme systems involved in the synthesis of cellular material in fungi. We know most about the glyoxalate by-pass enzymes, isocitratase and malate synthetase, with which the fungi can make four-carbon compounds from acetate. From malate, a series of conversions and decarboxylation will give pyruvate, a three-carbon fragment. The Embden-Meyerhof system can then continue its subsequent conversion to hexoses. Carbon dioxide fixation to form phosphoenolpyruvate, thence forming the oxaloacetate, has been shown [15]. Oxaloacetate can then either enter the Krebs cycle or be combined with ammonium ion to form aspartic acid. By a similar amination, α-ketoglutarate from the Krebs cycle can form glutamic acid. The glyoxalate pathway in fungal spores differs from that in bacteria in that it is constitutive, that is, formed when the spore is constituted; in bacteria it is adaptive, that is, produced by the cell only if acetate is present in the medium [14, 16]. If results from the vegetative cells of other fungi can be applied to spores of these mycelial forms, the spores should also contain enzymes for chitin synthesis.

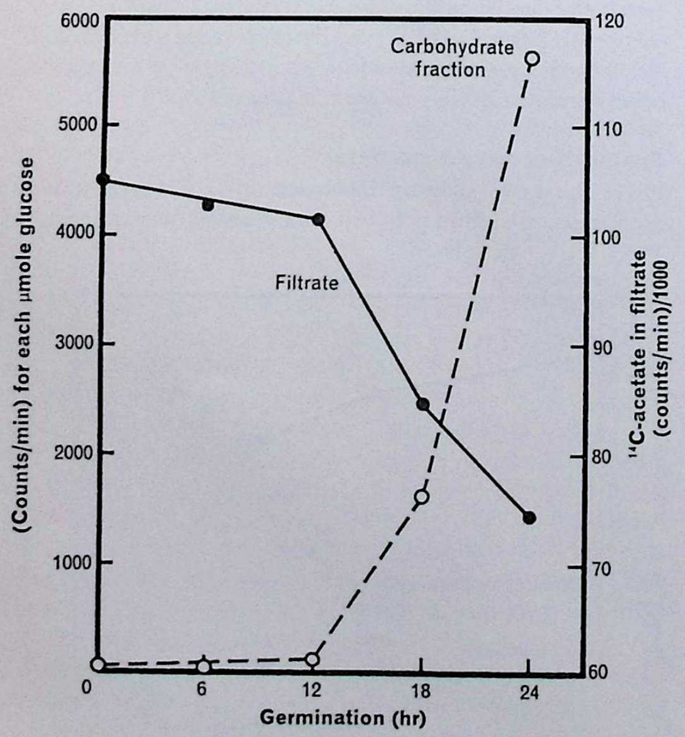


Figure 3 Incorporation of acetate-2-^{14}C into carbohydrate by uredospores of *Puccinia graminis tritici* after different periods of germination.

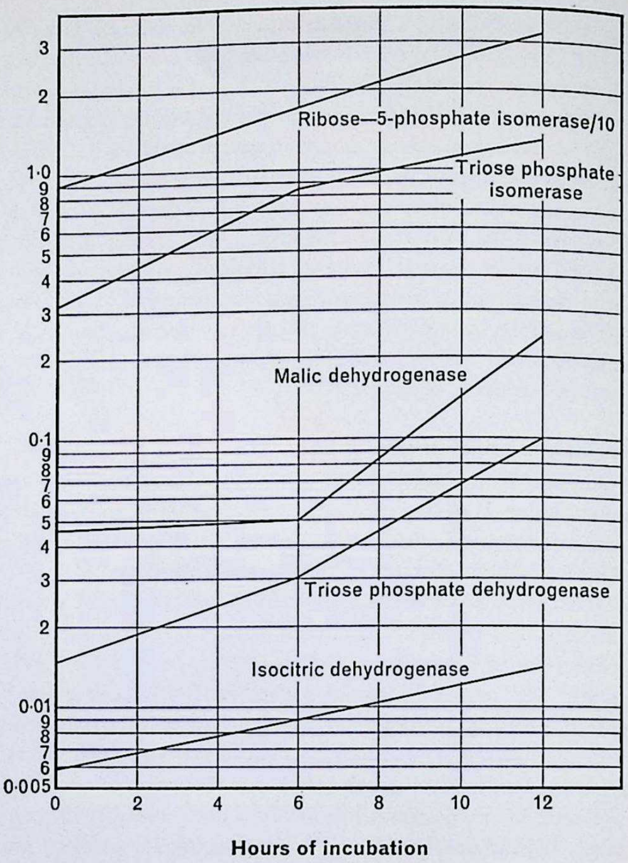


Figure 4 Synthesis of some enzymes during germination of *Ustilago maydis* before germ tubes were formed. Enzymes having no initial activity and detectable only after 12 hours are not shown. Specific activity is in μmoles substrate transformed per min per mg protein.

Though protein synthesis might be considered as including enzymes, special mention must be made of enzymes because of their metabolic importance. Three of the four fungi on which we made intensive studies for enzyme activity contained all the enzymes involved in normal glucose catabolism. Spores of the fourth fungus, *Ustilago maydis*, had no detectable activity for seven of the enzymes. One of these, α-glycerophosphate dehydrogenase, is usually associated with the Embden-Meyerhof system. Three others, glucose-6-phosphate dehydrogenase, transketolase, and transaldolase belong to the pentose cycle. The remainder, aconitase, fumarase, and succinic dehydrogenase, are part of the Krebs cycle. If one postulated, on the basis of the evidence previously discussed, that germination depended on prior metabolic activity, spores of *Ustilago maydis* should not have germinated at all. A study of enzyme activity at various periods during incubation revealed that the enzymes which could not be detected in the spore prior to its incubation were later synthesized [15]. In effect, all the enzymes required for energy production from glucose were made before germ tubes began to form. During this period, further synthesis was occurring of those enzymes that had been initially present. The increases varied from threefold to sevenfold according to the enzymes (figure 4). At the same time, there was a sharp rise in the incorporation of radioactive phenylalanine and in respiration (figure 5).

Energy and germination

The time required for germination among the different species of fungi might be related to the vigour of their metabolism. Of the species with a full complement of enzymes, the two rusts germinated within 2 hours and *P. oxalicum* within 6–8 hours.

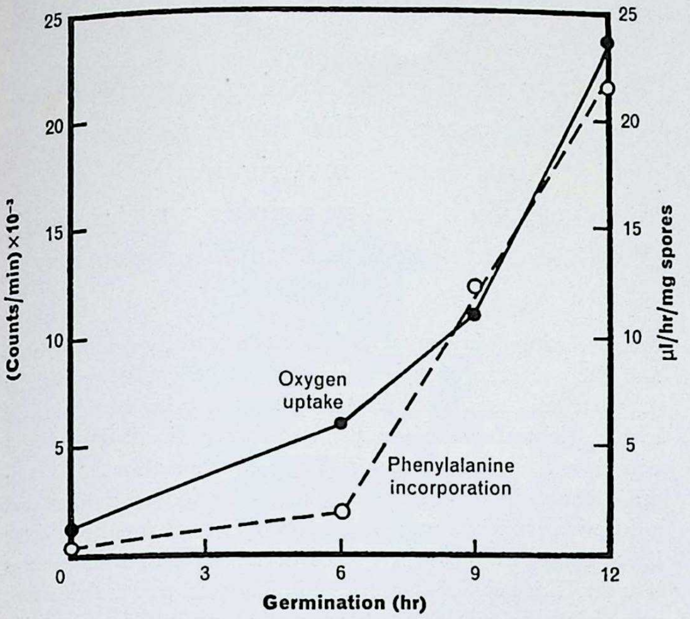


Figure 5 Incorporation of phenylalanine-1-C^{14} into protein and oxygen uptake at different stages of germination of telospores of *Ustilago maydis.*

Ustilago maydis, on the other hand, did not begin to germinate until after incubation for 12 hours, and even at that time only a slight deformation of the cells was observed. This long period corresponds to the time required for the formation of a full set of enzymes, and leads to the hypothesis that the formation of germ tubes is an energy-requiring process depending on the presence of substrate and of enzymes in sufficient concentration to produce the critical amount of energy necessary for the transformation of spore cells into germ tubes. Different fungi will, of course, have different mechanisms for producing the energy and will differ in the efficiency of their enzyme systems. Further studies to correlate enzyme synthesis with ability to produce germ tubes are needed before the hypothesis can be fully tested. The occasional requirement for vitamins, or in some mutants for some particular amino acid, to secure germination, can also be interpreted as part of the requirement for enzyme activity.

Though the role of energy-producing mechanisms has been emphasized, the synthesis of all kinds of metabolites is needed for balanced function in germination. Nevertheless, the release of energy appears at present to be the critical factor. Enough of the other metabolites are already incorporated in most spores to allow the machine to run if energy can be produced. Energy requirements might also be related to the breaking of dormancy. The heat activation studies of Goddard involved the activation of both anaerobic and aerobic systems of glucose degradation. Activation by heat initiated the anaerobic system, and germ-tube formation was correlated with the development of the aerobic system. Oxidative mechanisms were also postulated by G. L. Farkas [4] as being involved in the stimulation of germination of rust spores; the removal of self-inhibiting effects by the addition of various chemicals was accompanied by their ability to increase oxygen consumption. A comparison of the respiratory and synthetic activities of spores of saprophytes and obligate parasites by R. C. Staples and his colleagues [17, 18] showed that the rate of synthesis and respiration was higher in the saprophytes than in the obligate parasites. The results seemed to indicate that the limited ability to synthesize protein might explain, and be responsible for, obligate parasitism. One might well inquire why the rust fungi have complete systems for the breakdown of sugar, although they use fatty acids for their oxidizable carbon during germination. One explanation is that though these systems are not needed for germination, they are necessary once the fungus begins its parasitic existence on the host plant. The organism would not then need to synthesize further enzymes *de novo* but could begin its parasitic metabolism immediately on penetrating the host.

Redefinition of spore germination

It can now be seen that spore germination needs more precise definition. No longer can it be defined by morphological changes alone; it must also be related to functional changes. Germination is the process by which a spore is transformed from a dormant state, of low metabolic activity, to one of high activity. Formation of the germ tube is the outward and visible sign that this metabolic change is complete.

References

[1] Allen, P. J. *Plant Physiol.*, 32, 385, 1957.
[2] Forsyth, F. R. *Canad. J. Bot.*, 33, 363, 1955.
[3] Yarwood, C. E. *Mycologia*, 48, 20, 1956.
[4] Farkas, G. L. and Ledingham, G. A. *Canad. J. Microbiol.*, 5, 141, 1959.
[5] Goddard, D. R. *Cold. Spr. Harb. Symp. quant. Biol.*, 7, 362, 1939.
[6] Gottlieb, D. *Bot. Rev.*, 16, 229, 1950.
[7] Kornfield, J. M. and Knight, S. G. *Bact. Proc.*, 19, 151, 1960.
[8] Owens, R. G., Novotny, H. M., and Michels, M. *Contr. Boyce Thompson Inst.*, 19, 355, 1958.
[9] Kopellof, N. and Kopeloff, L. J. *Agric. Res.*, 18, 195, 1919.
[10] Yanagita, T. *Arch. Mikrobiol.*, 26, 329, 1957.
[11] Shepard, C. J. *J. Gen. Microbiol.*, 16, 1, 1957.
[12] Caltrider, P. G. and Gottlieb, D. *Phytopathology*, 53, 1021, 1963.
[13] Reisener, H., McConnel, W. B., and Ledingham, G. A. *Canad. J. Biochem. Physiol.*, 39, 1559, 1961.
[14] Caltrider, P. G., Ramachandran, S., and Gottlieb, D. *Phytopathology*, 53, 86, 1963.
[15] Staples, R. C. and Weinstein, L. H. *Contr. Boyce Thompson Inst.*, 20, 71, 1959.
[16] Gottlieb, D. and Ramachandran, S. *Mycologia*, 52, 559. 1960.
[17] Staples, R. C., Birchfield, H. P., and Baker, J. G. *Contr. Boyce Thompson Inst.*, 21, 97, 1961.
[18] Staples, R. C. Syamananda, R., Kao, V., and Black, R. J. *Ibid.*, 21, 345, 1961.

Field-ion microscopy

D. G. Brandon

Many of the weakening faults in crystals of a metal or alloy relate to single atoms rather than to groups, and are invisible under the electron microscope. The field-ion microscope shows single atoms and makes it possible to study these faults and to investigate, for example, surface reactions on the atomic scale. This article describes the principles that underlie the field-ion microscope and discusses both metallurgical and non-metallurgical applications.

One of the major problems of metallurgy has been, for a long while, that of relating gross changes in the mechanical properties of a metal or alloy to small differences in microstructure. The science of metallography, the microscopical examination of metals, has developed in response to this need. The ordinary optical microscope is limited to a resolution of about one thousand atom diameters, and the electron microscope has proved to be by far the most successful means of extending the range of microstructural investigation to higher resolutions. Over the past ten years, new methods have been developed for preparing thin films of a metal from bulk specimens for electron microscope examination. With these techniques, the microstructure can now be directly observed down to a resolution of a few atom diameters. However, while the electron microscope has revealed detail in thin metal films that was previously invisible, and often unsuspected, the resolution is not sufficient to reveal the atomic structure of the metal itself. It is this last step in the analysis of microstructure that can be made with the field-ion microscope.

The elucidation of the arrangement of atoms is not of academic interest only. Many of the changes in mechanical properties of metals and alloys are related to the presence of single atomic defects rather than clusters of defects. For this reason much of our knowledge of, for example, the nature of the interface between two crystals, has, until recently, been based on intelligent guesswork rather than on any experimental evidence. The field-ion microscope is of use here, and it also has applications in non-metallurgical studies. It gives a picture of the atomic structure of a metal surface. Little is known about the way in which chemical reactions take place on a clean surface, so that the field-ion microscope is valuable in that one can directly observe catalytic and corrosion reactions in atomic detail.

The field-ion microscope is the invention of one man, Erwin Müller: and his research group, at Pennsylvania State University, is still one of the more active in developing the technique and applying it to new problems. The microscope was originally developed by Müller in 1951 while he was working in Berlin on an earlier technique, field-emission microscopy, which employed basically the same specimen geometry. The technique is briefly discussed later in this article. These preliminary results were extended and finally published in 1956, by which time Müller had moved to the United States, but it was not until 1958 that the potentialities of the technique became fully apparent. In that year, at the Fourth International Conference on Electron Microscopy in Berlin, Müller was able to show micrographs from a large number of metals in which the regularity of the crystal lattice was beautifully demonstrated and the presence of defects in this lattice could be readily detected [1]. Since then, field-ion microscopy has been applied to a large number of research problems, including grain-boundary structure, radiation damage, order-disorder phenomena, the structure of solid solutions, oxidation behaviour, and adsorption processes [2, 3]. Much of this work has been done in the United States, although one of the larger research groups is at the University of Cambridge in England and a significant proportion of the work has been carried out in Germany.

Principles of the microscope

The basic microscope assembly is shown in figure 1. A small amount of a pure inert gas, usually helium, is admitted into the vacuum chamber and, a positive potential is applied to the fine wire specimen. The gas is attracted towards the specimen point by polarization forces. If the specimen wire has a sufficiently small end radius, the electric field at its tip will

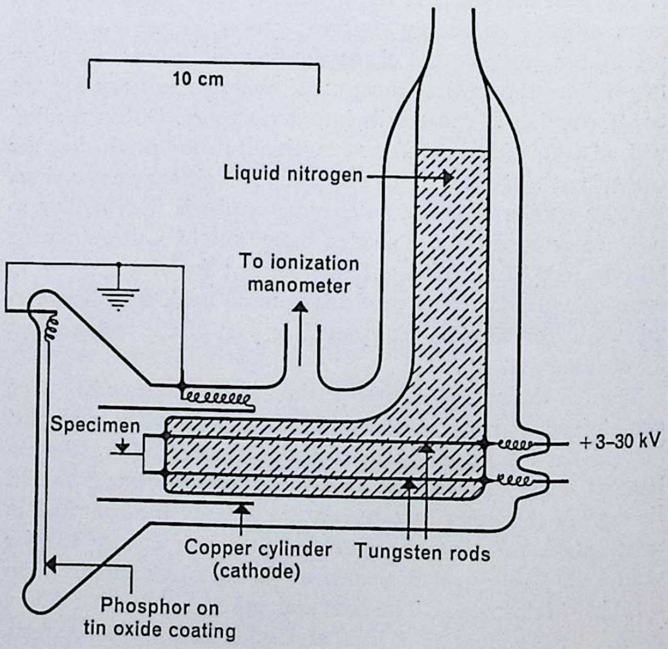


Figure 1 The field-ion microscope.

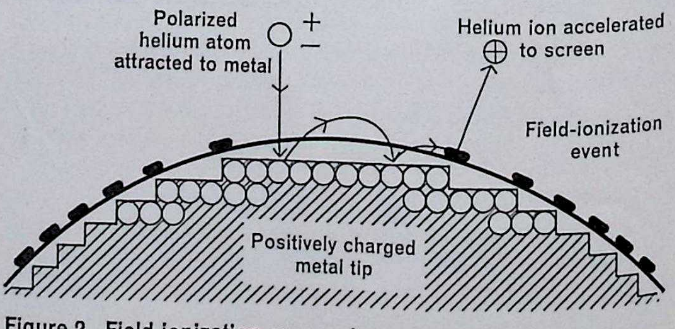
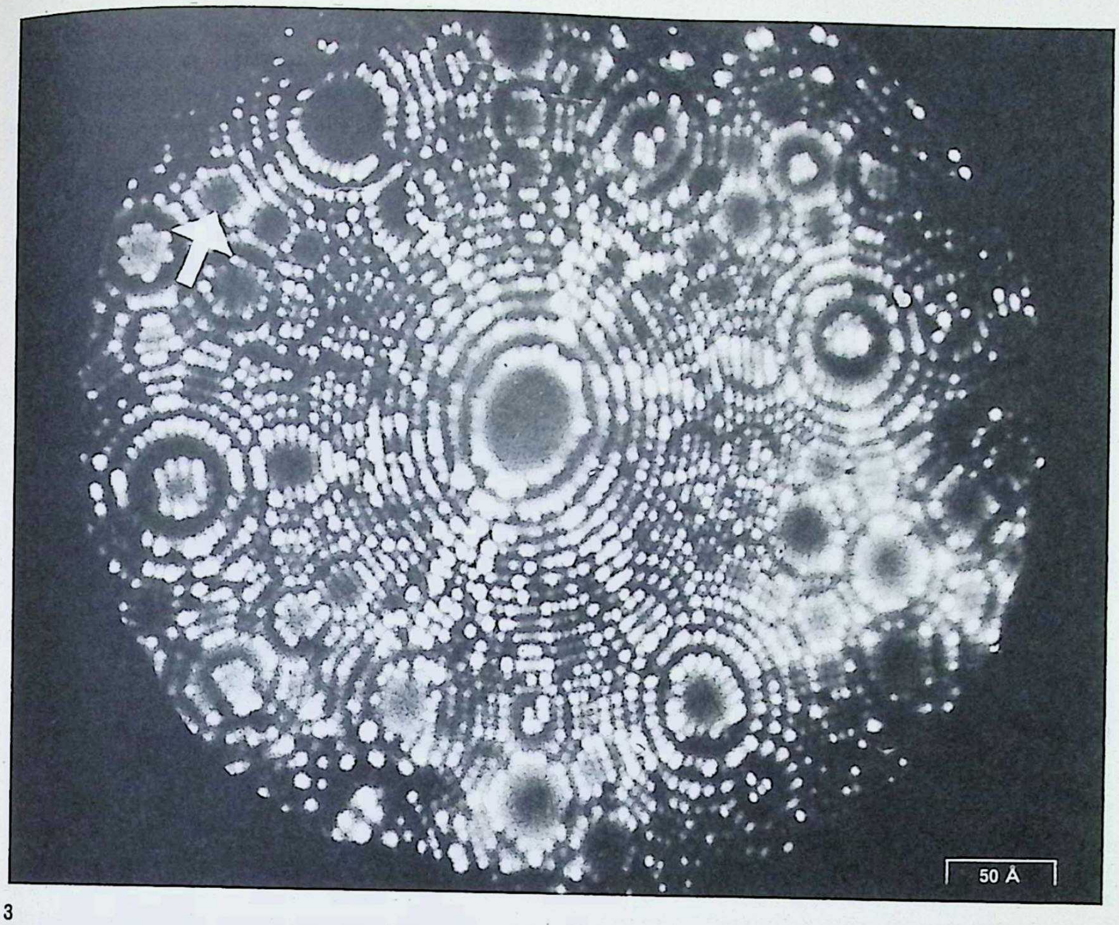

Polarized helium atom attracted to metal
Helium ion accelerated to screen
Field-ionization event
Positively charged metal tip

Figure 2 Field ionization over a charged metal surface.

D. G. Brandon, *M.A., Ph.D.*
Was born in London in 1935 and educated at the University of Cambridge. He remained at the Department of Metallurgy, Cambridge, after taking his doctorate and initiated research on field-ion microscopy there, finally becoming Senior Assistant in Research to the Department. He is now at the Battelle Memorial Institute, Geneva, where he is applying the field-ion microscope to the study of steels and other non-refractory metals.


3

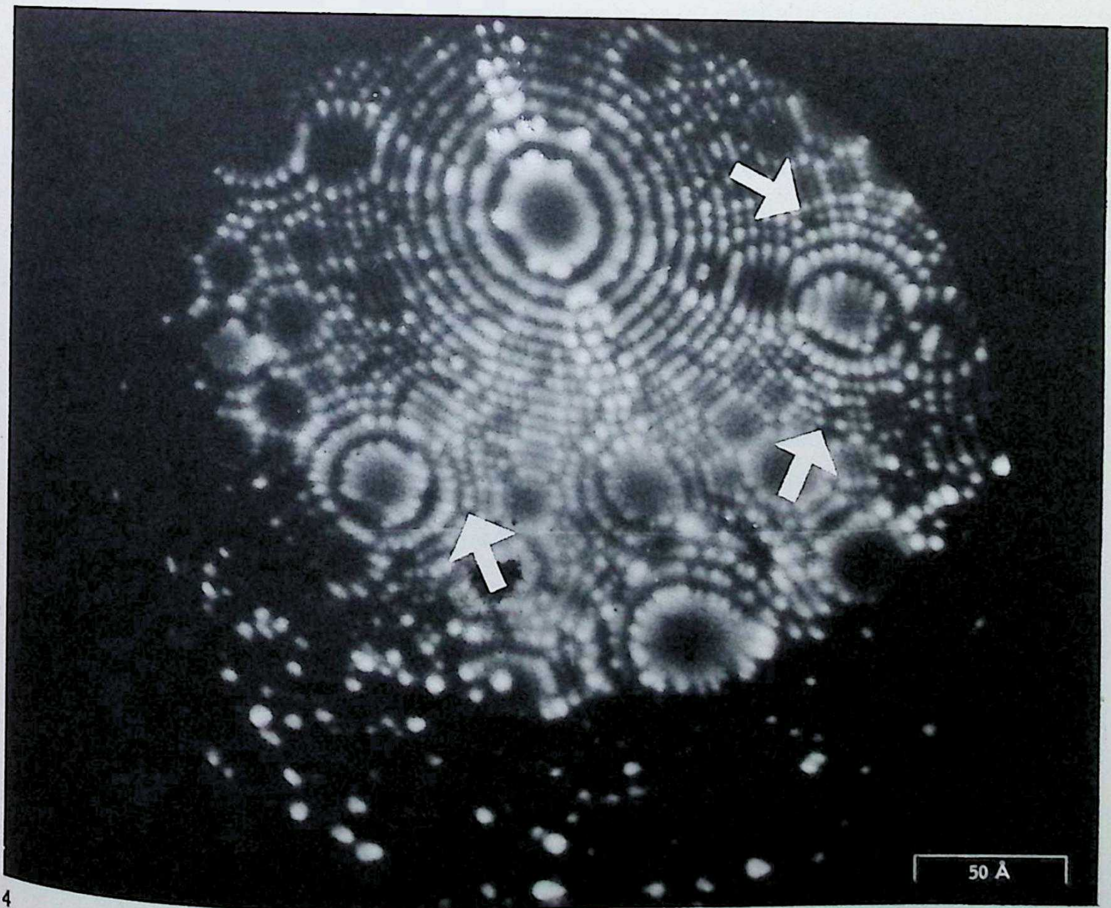

4

Figure 3 Helium field-ion micrograph of a tungsten point. A vacant lattice site is arrowed. The axis of the tip is tilted towards the top of the micrograph. Other surface irregularities are due to irregular field-evaporation near the tip shank (dark area at the bottom of the micrograph).

Figure 4 Vacancies and a vacancy cluster in reactor-irradiated tungsten.

be large enough to ionize the helium gas in this region (figure 2). The ions are then accelerated radially outwards along the lines of force and impinge on a fluorescent screen. Here they form a projection image of the end of the wire. Specimens are prepared by electrolytically polishing the wire to an end-radius of about one hundred atom diameters. Since the magnification is approximately the ratio of the point to screen distance to this end-radius, a magnification of one million times is quite easily achieved (1 atom diameter $\simeq 3 \times 10^{-8}$ cm). The probability of ionization depends sensitively on the local atomic structure of the surface of the point, and the image therefore reflects this atomic structure (figure 3). Provided the temperature is sufficiently low (less than 50°K) to keep the thermal velocity of the gas small, the resolution depends only on the radius of the ions used to form the image. The electric field strength required to produce ionization depends principally on the ionization potential of the gas and the working temperature. With a gas of high ionization potential, very high electric field strengths are required (about 3 volts/Å), and only the refractory metals can withstand such fields without themselves being ionized and evaporated. The working range of the microscope for a particular combination of gas, metal, and temperature is defined as the ratio of the evaporation field for the metal to the threshold field for obtaining an image [3] (figure 5). Since this working range can be increased by using a gas of lower ionization potential, it might be asked why helium, with the highest ionization potential of any element, is the preferred image gas. There are several good reasons for this. Firstly, the size of the helium ion is small ($\simeq 2$ Å) and the resolution correspondingly good. Secondly, helium is not adsorbed on a metal surface at temperatures much greater than 5°K and consequently does not interfere with the surface itself. Thirdly, at the high electric fields needed to ionize helium, all chemically active gases in the vacuum system are ionized long before they can reach the metal point. This removes the need for the extreme purification of the residual gas that is normally needed for work on clean metal surfaces. Finally, helium has a relatively high phosphor-efficiency, making it possible to obtain a visible image, even at very low ion-currents, without the use of an image intensifier (see the scale on figure 5).

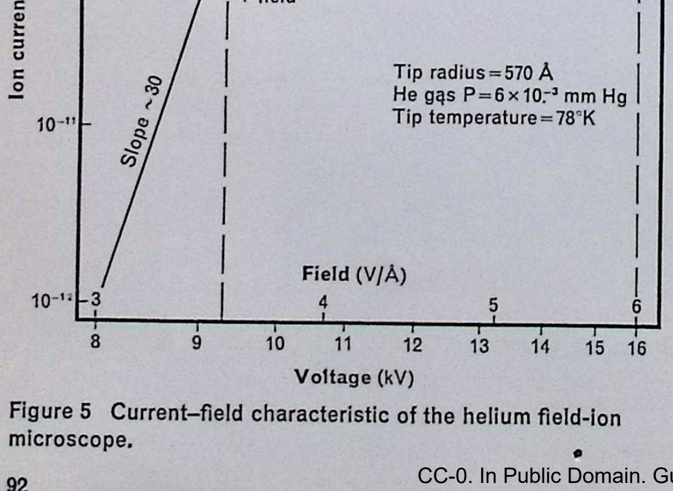


Figure 5 Current–field characteristic of the helium field-ion microscope.

Most of the limitations of the field-ion microscope are related directly to the above factors. The evaporation field of a metal is nearly constant at low temperatures and is dependent only on the ionization potential, work function, and sublimation energy of the metal. Of these, the sublimation energy is usually the controlling factor. This is why only the more refractory transition metals will form an image with helium.

The threshold field for neon is about 40 per cent lower than for helium. This would make it more useful, but the reduced phosphor efficiency, increased gas-scattering of the ions, and increased sensitivity to contamination, all present new experimental difficulties. Each can be dealt with. The reduction in phosphor efficiency can best be overcome by using either an image intensifier, or image converter [4] that will convert the ion image to an electron image: the increased gas-scattering can be prevented by reducing the working pressure, and hence the available ion-current. The only way to prevent the effect of contamination is to remove it. This involves correspondingly more stringent vacuum requirements, which means that it takes longer to change a specimen. An additional complication is that one has to bake the vacuum assembly to get rid of adsorbed contamination on the walls of the system.

Image gases with a lower threshold field than neon—argon, for example—have not been much used, possibly because of the very much more stringent vacuum requirements, and because of the larger diameter, and consequently poorer resolution, of the gas ions. The introduction of an image converter rather than an image intensifier seems to be the best means of improving the performance of an argon-ion microscope. Argon-ion microscopy is at present limited by the very low phosphor efficiency for argon, rather than by more fundamental disadvantages, since ultra-high-vacuum systems are becoming increasingly easy to build and maintain, and most workers would willingly make some sacrifice of resolution if a stable image of the non-refractory metals could be obtained.

The observed reduction in threshold field with temperature has not yet been fully exploited. It is known that helium is adsorbed on a metal surface at 4·2°K, the boiling point of liquid helium. It is obviously not possible to work below this temperature. However, the lowest temperature that can be used is not known. The minimum temperature used at present is 21°K (the boiling point of liquid hydrogen). Extrapolation of the known data to 5°K suggests that very much more stable images could be obtained at this lower temperature. Stable helium images of iron and nickel are only obtained with difficulty at 21°K, so that there are obvious advantages in lower temperatures of operation.

Other difficulties of the technique are less obvious. Firstly, the field of view is limited and, on the atomic scale, defects are few and far between, so that it requires careful examination of many micrographs to spot them. The field can be used to evaporate the specimen atom-layer by atom-layer, so as to obtain a three-dimensional picture of the structure, but the total volume available for examination is still truly minute ($\simeq 10^{-15}$ cm^3). Again, only point specimens can be examined, and it is difficult to prepare specimens from anything but fine wire. Still more important, the image is always obtained at an electric field close to that for field-evaporation, and, in any case, field-evaporation is always used to clean and smooth the surface. The electrostatic stresses induced in the specimen at these fields are enormous, and many metals can be seen to deform plastically while under observation. Finally, only the electric field distribution over the surface is observed. This means that any image detail seen can be related to the sub-surface structure only if the effects at the surface are known and a correlation is

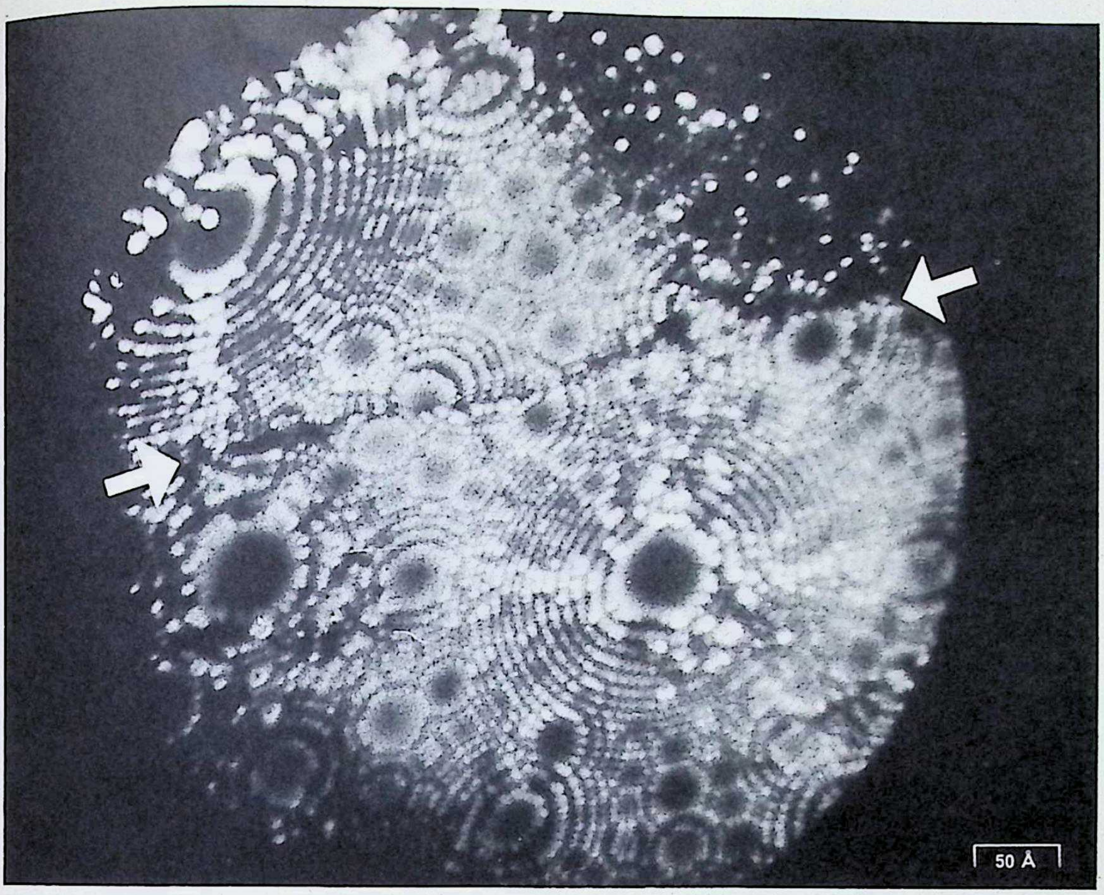


Figure 6 A grain boundary in tungsten. The dark regions (for example at top right) are surface irregularities near the tip shank and have no microstructural significance.

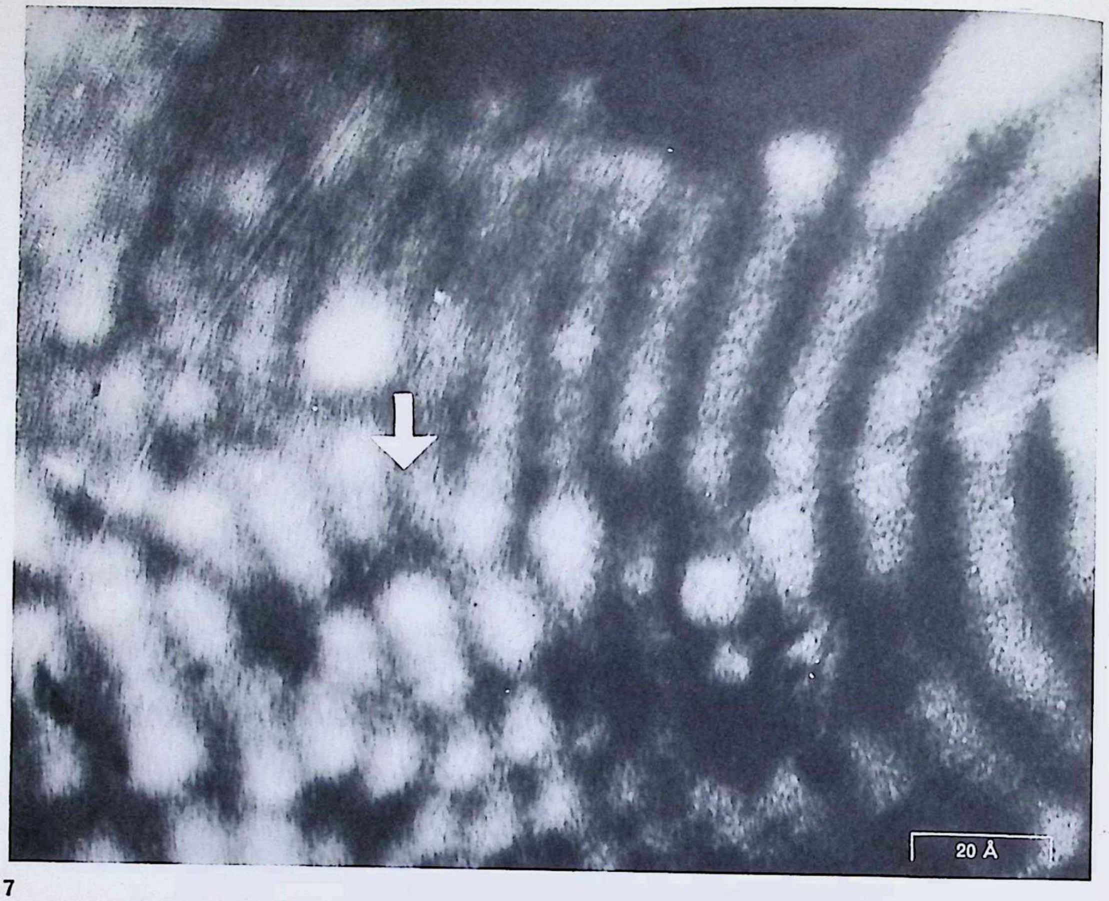
established. In practice this can often be done, but there are instances where the precise surface structure is unknown and the interpretation of the image is doubtful.

With such a formidable list of limitations and disadvantages it may seem somewhat difficult to account for the growing use of the field-ion microscope and its rapidly expanding range of applications. However, the limitations are known and understood, and the theory of image formation is for the time being well in advance of experimental developments. The direction these developments will take now seems quite clear, and the future applications of the microscope are well defined. At the same time the remarkable resolution of the technique ensures that the information obtained will be unique. The field-ion microscope has already made a significant contribution to metallurgical science that can readily be demonstrated by a few examples. Those given below are taken from the work of the Cambridge group, and include studies of radiation damage, grain boundaries, dislocations, and solid solutions.

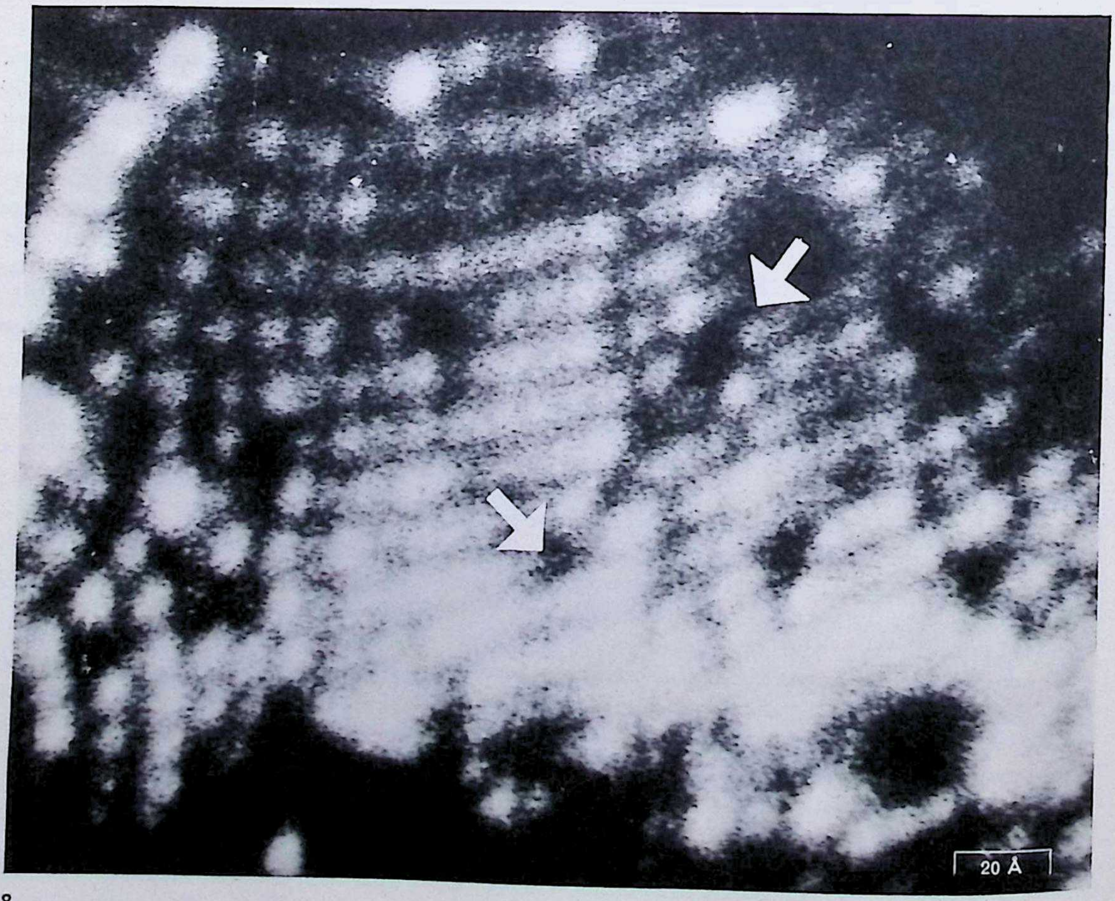
Applications

Radiation damage. The increasing use of nuclear power has highlighted the problem of 'radiation damage' in metals. This is defined as the deterioration in the mechanical properties resulting from bombardment with high-energy particles. It is known that radiation damage occurs when metal atoms are knocked out of their normal positions in the lattice to produce vacant lattice-sites and extra 'interstitial' atoms. It is also known that the most serious effects of radiation damage are felt when the extra atoms and vacant sites form small clusters. However, in just that range of cluster size that is of greatest interest, the electron microscope fails to resolve the structural details of the clusters. Indeed, in many cases, notably that of steel, the electron microscope shows no apparent changes in the microstructure until the mechanical properties have been catastrophically affected. The field-ion microscope, on the other hand, is capable of resolving even the individual vacant lattice sites (figure 3), and not only the presence but also the detailed morphology of the clusters can be seen (figure 4). Two types of radiation-damage experiment have been performed at Cambridge [3]. Firstly, the field-ion microscope has been used to study in atomic detail the microstructure of specimens prepared from reactor-irradiated material. The vacancies and the vacancy clusters seen in figure 4 were produced in the bulk metal by reactor irradiation. The image was produced by field-evaporation of the metal in the microscope. Another method is to irradiate the specimen in the microscope itself and observe the damage at the surface as it occurs. In this way the surface damage produced by single energetic particles has been photographed [3].

Grain boundaries. Where two crystal lattices of the same metal but different crystallographic orientations meet, there is a narrow region of disorder. This has been known for many years. What was not known was the extent of this disordered region, its exact nature, and its dependence on the misorientation across the interface. Field-ion microscope observations showed that this boundary region was extremely thin and not very homogeneous (figure 6). There are regions of good fit in which the two crystal lattices seem to be completely contiguous, with no transition region, but there are also regions where the fit is not so good, and where a transition region two or three atoms in diameter exists between the two crystals. The extent of the regions of good fit depends on the exact type of boundary, and, from the field-ion microscope observations, it has been possible to formulate a theory of grain-boundary structure which accounts for these microstructural observations and also for many other hitherto poorly understood properties



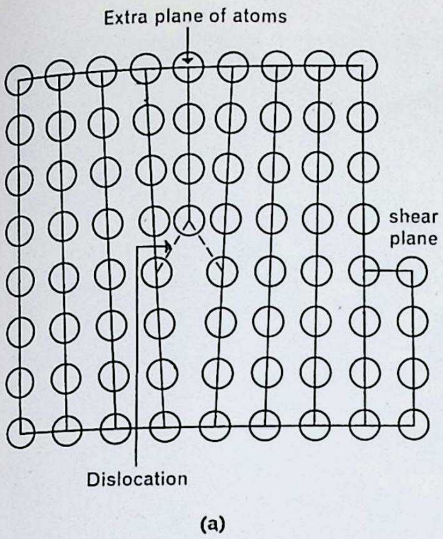
7



8

Figure 7 An edge dislocation in molybdenum.

Figure 8 A partial dislocation and stacking fault in an ordered cobalt-platinum alloy.



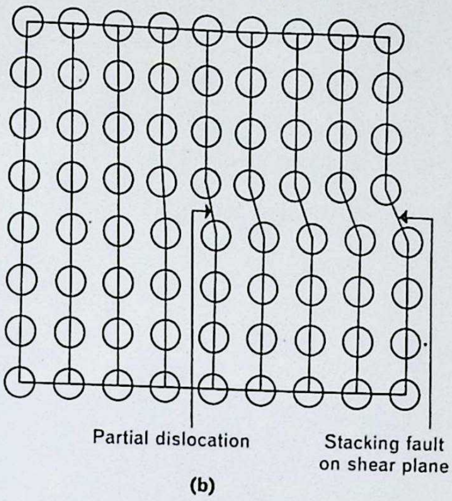


Figure 9a Edge dislocation in a simple cubic lattice.

Figure 9b A partial dislocation and stacking fault in a simple cubic lattice.

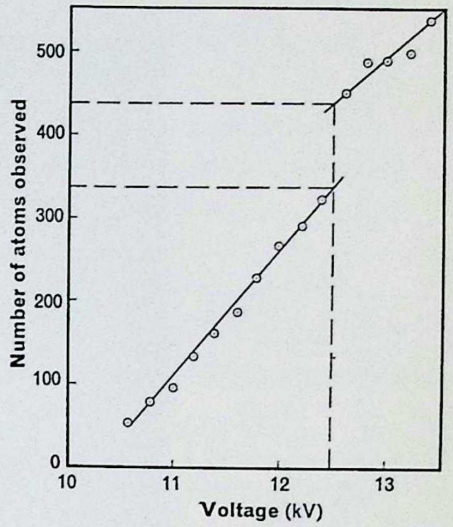


Figure 10 Number of imaged atoms as a function of voltage for the tungsten 26 per cent rhenium alloy.

of boundaries. So far this investigation of boundary structure has been confined to tungsten and to dilute tungsten-rhenium alloys, but there is no reason why the work should not be extended to other metals. One of the more interesting observations made during this study of grain boundaries was that of the presence of irregularly placed impurity atoms at the boundary. Premature fracture of a metal in service is often caused by segregation of impurities at the grain boundaries, and there now seems every reason to suppose that the field-ion microscope will help us to understand some of the factors responsible for this behaviour.

Dislocations. The microstructural features responsible for plastic flow are the line defects, or 'dislocations', in the metal lattice. The simplest type of dislocation, the 'edge' dislocation, may be thought of as an extra half-plane of atoms inserted in the crystal. Movement of such a dislocation through the crystal results in the shear of one half of the crystal over the other half by one complete atomic distance (figure 9*a*). The ease with which a dislocation can move through the crystal depends critically on the binding of the atoms at the core of the dislocation, and the atomic structure of this core can be resolved in the field-ion microscope (figure 7). In many metals there are 'partial' dislocations which, instead of shearing one half of the crystal over the other by one complete atomic distance, leave the two halves of the crystal in a metastable intermediate position separated by a 'stacking fault' (figure 9*b*). These faults are visible in the field-ion microscope as discontinuities in the atomic planes. An example of a partial dislocation and the stacking fault left behind it is shown in figure 8. This fault is in a completely ordered magnetic platinum-cobalt alloy, in which each type of atom occupies a specific type of lattice site. In the same alloy it has been possible to follow the transition from the disordered state, in which the atoms are randomly placed, to the ordered state, and it was observed that large numbers of vacant lattice sites and dislocations were introduced during ordering. The maximum hardness of the alloy, its optimum magnetic properties, and the greatest density of defects were all found to occur in the same specimen after the same heat treatment.

Analysis of special alloys. With a random solid solution containing more than 5 per cent, in atoms, of the alloying element, extremely complicated field-ion images have been obtained. Instead of the striking regularity found in the pure metal (figure 3), the image appears as a random array of bright spots of varying intensity (figure 11). The reasons for this difference are twofold. Firstly, the difference in atomic species is usually sufficient to ensure that one or other element is preferentially evaporated from the surface during the preliminary field evaporation process. While this will to some extent increase the randomness of the surface, it cannot account completely for the irregularity of the image. The second reason lies in the localized differences in electronic structure. These ensure that even two atoms in similar geometric positions in the lattice will give different image intensities if they are of different atomic species. This effect is seen most clearly if the number of atoms whose image is formed is plotted as a function of the average voltage (figure 10). The jump in the curve is directly proportional to the atomic concentration of the alloying element. Although the detailed reasons for this effect are by no means understood, the possible applications of this type of 'voltage analysis' are numerous.

Applications outside metallurgy

So far only the metallographic potentialities of the field-ion microscope have been considered, but the technique has also been applied with great success to field-emission and mass-spectrometric problems. In field-emission microscopy, a technique that dates from 1936, the specimen geometry is identical to that used in the field-ion microscope, but with a negative specimen-potential. Electrons are drawn from the metal surface and accelerated radially on to the fluorescent screen by an electric field of an order of magnitude less than that employed in helium field-ion microscopy. The projection electron image obtained has a relatively poor resolution (≃ 10 atom diameters), but it accurately mirrors the variations in electron potential from place to place on the metal surface. Since adsorbed layers on the surface strongly affect this electron potential (the 'work function' of the surface), the field-emission microscope can be used to follow the kinetics of adsorption and investigate any orientation dependence of the adsorption process. However, while a great deal of information on both chemical and physical adsorption has been obtained in this way, the metal surfaces used have been poorly characterized and the details of the process have not always been clearly understood. By using the same specimen for both field-emission and field-ion microscopy, the microstructural features of adsorption

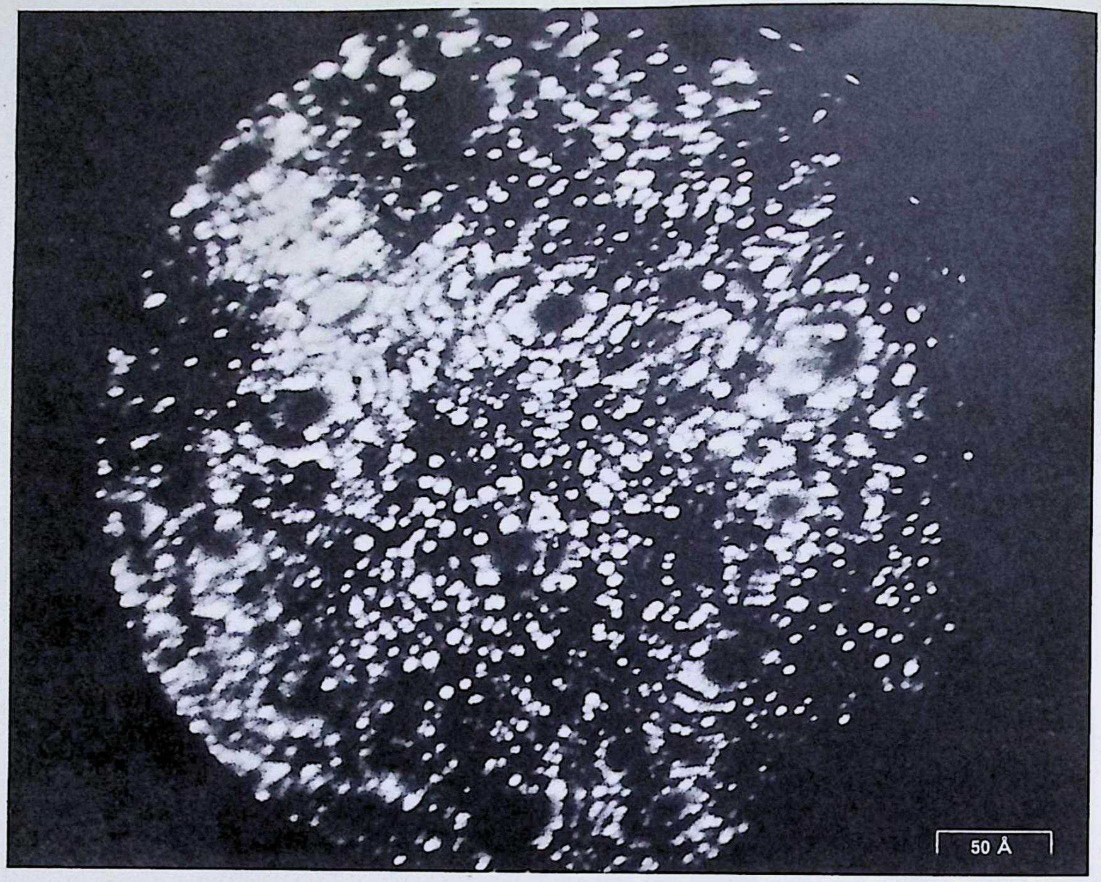


Figure 11 Helium field-ion micrograph of a random solid solution: 26 atomic per cent rhenium in tungsten. The atomically rough surface results from the preferential evaporation of rhenium from the lattice.

can be followed on a specimen surface characterized in atomic detail. This technique has been developed and used by Gert Ehrlich at the General Electric Research Laboratory in America to follow the chemical adsorption of numerous inorganic gases [5]. Using the same technique it should be possible to investigate the factors controlling catalytic activity on metal surfaces.

Because field-ionization is sensitively dependent on the electric field and has a sharply defined threshold field (figure 5), the ions produced by this process are effectively mono-energetic. This is true for the inert gases used to form a field-ion image, and also for quite complicated organic compounds. Also, whereas thermal ionization, or ionization of large organic molecules by electron bombardment, gives a large number of ionized fragments, field ionization gives only a few large fragments corresponding to breakdown of the molecules at their weakest point [6]. Field-ion microscopes are therefore being increasingly used as ion sources for mass-spectrometric analysis. The metal specimen now serves as a sink for the electrons extracted during ionization and in most cases the compounds being analysed have very low ionization potentials. It is important that the metal surface should not be chemically attacked by adsorbed gases, and gold and platinum are therefore the preferred metals for use as ion sources. The sensitivity of this combined mass-spectrometer and field-ion microscope can be of the order of 10^{-18} amps (6 ions/second), and it is theoretically possible to analyse the metal specimen itself by field evaporation. That this has not yet been done is largely because of the large number of other possible developments, but there is no doubt that eventually it will be done and that 'counting the atoms' as they evaporate will prove to be a most sensitive and accurate method of micro-analysis for metals and alloys.

In this article the emphasis has been not so much on what has already been done with the field-ion microscope as on what could be done with it in the future. The technique is still very new and, as with all new techniques, there has been an intervening period between development and application during which it has been regarded with considerable scepticism. To be able to see atoms is a striking achievement, but it cannot by itself justify the introduction of the microscope as a standard research tool. It is only by the production of unique information that can contribute to our understanding of the properties of metals and metal surfaces that the field-ion microscope may be judged successful. On the showing of the past few years it is very successful indeed.

Acknowledgments

I should like to record my thanks to all my former colleagues at Cambridge, and especially to Dr M. J. Southon and Mr B. Ralph, who were responsible for much of the experimental work reported here. I should also like to thank *The Times Science Review*, the *Journal of the Physical Society of Japan*, the *British Journal of Applied Physics* and the *Philosophical Magazine* for permission to reprint micrographs.

References

[1] Müller, E. W. '4th International Conference on Electron Microscopy', p. 820. Springer, Berlin. 1958.
[2] *Idem*. International Conference on Crystal Lattice Defects, Kyoto. 1962. *J. phys. Soc. Japan*, 18 (Suppl. II), 1, 1963.
[3] Brandon, D. G., Wald, M., Southon, M. J., and Ralph, B. International Conference on Crystal Lattice Defects, Kyoto. 1962. *J. phys. Soc. Japan*, 18, (Suppl. II), 324, 1963.
[4] Brandon, D. G. *Brit. J. appl. Phys.*, 14, 474, 1963.
[5] Ehrlich, G. and Hudda, F. G. *J. chem. Phys.*, 36, 3233, 1962.
[6] Beckey, H. D. *Z. Naturforsch.*, 17a, 1103, 1962.

The relief of pain—the search for the ideal analgesic

K. W. Bentley

The relief of pain is one of the more important of all medical objectives. For centuries, morphine has been pre-eminent as an analgesic, but its side-effects, especially on respiration, make it a far from ideal drug. Exhaustive investigations of substances of the morphine type strongly suggest that the undesirable side-effects are inseparably linked with the analgesic properties. The chance discovery that nalorphine, a morphine antagonistic, is a potent analgesic in man, though apparently not in experimental animals, has, however, opened up a promising new line of investigation. A feature of recent studies is the synthesis of compounds several thousand times as potent as morphine.

For many centuries the opium poppy, *Papaver somniferum*, has held a unique place in the armoury of the physician on account of the ability of the dried exudate of the cut seed capsule to relieve even severe pain without undue stupefaction. Thomas Sydenham (1624–89), often referred to as the English Hippocrates, maintained that 'few would be willing to practise medicine without opium'. Just over 150 years ago W. Sertürner [1] described the isolation of the active component of opium, morphine, so named after Morpheus, the god of sleep. This subsequently became generally available in a pure form to the medical profession as a result of the work of William Gregory [2]. The elucidation of the structure of morphine and the closely related bases occupied the attention of many eminent chemists for almost half a century and culminated in the work of Sir Robert Robinson and the late J. M. Gulland in 1925 and finally in the total synthesis of the alkaloid by Marshall Gates and G. Tschudi in 1952. During the course of this work so many fascinating molecular rearrangements were uncovered that Robinson was led to refer to the bases of this group as 'the star performers in the field of molecular acrobats'. The alkaloid was first commercially extracted by Gregory's process in 1833 by J. F. Macfarlan & Co. of Edinburgh, in whose laboratories the author has recently been engaged on some of the work described below.

Morphine still holds first place in the world as a major painkiller, but it is by no means the ideal drug. It has two serious disadvantages, in that it causes marked respiratory depression and very quickly renders the body physically dependent on its continued administration. Respiratory depression following the use of morphine in obstetrics is occasionally severe enough to produce foetal distress, which has to be reversed by the administration of nalorphine (see below), a specific narcotic antagonist. Morphine induces a state of profound mental relaxation and relief of tension, during which troubles are forgotten and the world seems a happy place, and it is for this reason that the unhappy, inadequate, or weak person initially turns to the drug. Unfortunately, as the effects wear off the euphoric state is succeeded by one of depression and gloom, from which release is gained by a further dose of the narcotic. In this way what starts as a psychologically induced habituation rapidly becomes a true physical dependence on the drug. Withdrawal of the narcotic is rapidly followed by the most distressing physical symptoms, almost immediately relieved by further administration of morphine or a morphine-like compound. Despite strict control, addiction to morphine and related drugs remains a great social evil. In medicine there is always the danger that the prolonged administration of morphine to a patient suffering chronic pain will result in the addiction of that patient, particularly as tolerance to the effects of morphine rapidly develops and larger and larger doses of the drug are necessary for the achievement of a constant degree of pain relief.

Over the last thirty years [3–7], an intensive search has been made for a new analgesic devoid of these two undesired side-effects. In particular, it was hoped to find a wholly synthetic drug, so that the cultivation of *Papaver somniferum* and the extraction of opium could be banned by international agreement. However, although the opium poppy is the most common source of narcotics, it is difficult to see how the prohibition of its growth would materially ease the solution of the narcotics addiction problem in civilized countries, especially in the United States. The illicit traffic is highly organized, and the relatively large-scale synthesis of pethidine, methadone, or one of the newer more potent substances should present no more difficulties than the extraction of morphine from opium, for the technical information necessary for the preparation of these highly addictive drugs is freely available in the literature. Indeed, addiction to the newer synthetic drugs is steadily replacing that to morphine and heroin in those countries in which they are available.

Evaluation of analgesic properties

The synthesis of new potential analgesics must be accompanied by the pharmacological appraisal of the compounds obtained in order to ensure that only those that are both effective painkillers [8] and relatively non-toxic are selected for further study in humans. It is probable that this testing constitutes the weak link in the chain from laboratory to hospital. Animals can be made to give a reproducible response to certain stimuli of a kind that are known to produce painful sensations in humans, and this response is abolished by morphine and other potent analgesics. Such stimuli include pressure on the rat's tail; heat on the mouse's foot or back; and electrical stimulation of the teeth. The animals' response usually takes the form of a flight reaction or a squeak, which may be interpreted as indicating that the animal is suffering pain. The tests most frequently applied in the pharmacological screening of analgesics are the tail-pressure test with rats and the hot-plate test with mice. In the former, analgesia is indicated by abolition of the squeaking response to tail pressure, and in the latter by disappearance of the flight response when standing on a hot plate. It is questionable, however, whether these stimuli produce anything resembling pathological pain in human beings, the relief of which is the whole object of the investigation. It is significant

K. W. Bentley, *M.A., B.Sc., D.Phil.*
Was born in Bradford in 1925 and educated at Bradford Grammar School and at Oxford University, where he was an ICI Fellow. Subsequently he was Lecturer in Chemistry at Aberdeen (1952–60) and Research Fellow in Organic Chemistry at Harvard (1956–57). Since 1960 he has been chief research chemist with the Joint Association of J. F. Macfarlan & Co. Ltd. and Reckitt & Sons Ltd. His research interests have always been in alkaloids, especially those of the morphine group, about which he has written extensively.

that they have in fact led only to the selection of a vast number of compounds possessing in general all the disadvantages of morphine itself, and clearly they represent a most unsatisfactory screening procedure.

The deficiencies of such tests are further emphasized by the fact that certain members of the series of morphine antagonists discussed below, which are active as analgesics in humans and provide the most promising area for further study, are inactive in both kinds of test. A better screening test, in which conditions are produced more closely akin to those under which the drug would be used, is obviously required. A promising method is the study of the abolition of the writhing or stretching response produced in certain animals by the intraperitoneal injection of phenyl-*p*-benzoquinone [9, 10]. The writer's colleague, R. E. Lister, has demonstrated that this writhing response is abolished by morphine-like drugs and also by many members of the morphine-antagonist group. He has established a correlation between addictivity and activity in the tail-pressure and hot-plate tests, but not between addictivity and the ability of the drug only to abolish this writhing response to phenyl-*p*-benzoquinone [11, 12].

Analgesic activity and chemical structure

The mode of action of morphine and similar compounds in relieving pain is not understood. It has been found that all potent analgesics have certain structural factors in common. All contain an aromatic nucleus linked to a carbon or non-basic nitrogen atom bearing no hydrogen and further linked through two carbons to a tertiary, or in rare cases secondary, basic nitrogen [4, 6]. The mirror image of morphine, (+)-morphine, is inactive, and all the synthetic analgesics have been found to be structurally related to (−)-morphine, their mirror images having greatly reduced activities. These facts imply that for analgesic action certain spatial requirements must be met and suggest that three-dimensional accommodation of the molecules at some receptor surface in the central nervous system is an essential part of the process by which the compounds exert their action. A. H. Beckett [13] has described the shape and approximate dimensions of such a hypothetical surface and shown how the major analgesics, but not their mirror images, could be accommodated there. What happens to the compounds at the receptor, however, is even less certain, though the remarkable difference between morphine and N-allylnormorphine, mentioned below, suggests that N-de-alkylation is an important step [14]. However, recent work by H. Rapoport and his co-workers [15] has provided no confirmatory evidence of this. It may be noted that normorphine, the secondary base, is generally of low activity when administered directly to the stomach of most animals, but is as active as morphine when administered into the cerebrospinal fluid [14, 16, 17]; its inactivity by the former route may be due to the difficulty of its passage through the blood-brain barrier.

Analgesics other than morphine

The work of preparing new major analgesics has proceeded along four distinct lines:

(i) Modification of the morphine molecule.
(ii) Synthesis of simpler substances, the molecules of which represent certain portions of the morphine molecule.
(iii) Elaboration of the morphine structure to provide more complex, more rigid molecules differing appreciably from the alkaloid in peripheral shape.
(iv) Synthesis of morphine antagonists of the N-allylnormorphine type from the above series.

Modification of the morphine molecule [3, 4]. Morphine (i) contains a number of modifiable centres, namely a phenolic hydroxyl, an alcoholic hydroxyl, a cyclic tertiary N-methyl group, a double bond, and a cyclic ether system. The effects on the analgesic power of changes at these centres are well known.

Esterification of the two hydroxyl groups marginally increases the activity, but diacetylmorphine (heroin) appears to be more euphoric than the parent alkaloid, and its addiction potential is accordingly regarded as being higher. Methylation of the alcoholic group also brings about some increase in activity, but alkylation of the phenolic hydroxyl greatly reduces the activity, codeine being only about one-tenth as active as morphine. Reduction of the double bond generally results in a two- or three-fold increase in activity, but in the case of the acetyl esters there is an appreciable decrease. Oxidation of the secondary alcoholic group to carbonyl, and replacement by either hydrogen or chlorine, all result in marked increases in activity, though the chloro-compounds are also substantially more toxic.

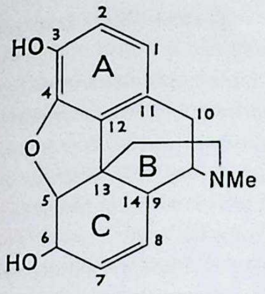
The conversion of the tertiary bases to quaternary salts, the opening of the nitrogen-containing ring by Hofmann degradation, and of the cyclic ether system by reduction, all result in almost complete loss of activity. The inactivity of the quaternary salts may, however, be due to their inability, as salts, to penetrate barriers on the way to the receptor surface.

The introduction of substituents on the alicyclic ring C has a variable effect. For example, 14-acetoxydihydrocodeinone (ii) and 5-methyldihydromorphinone (iii) (metopon) are about twelve times as potent as morphine, but 6-methylmorphine is only about half as active as the parent alkaloid.

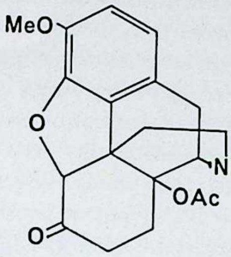
The treatment of heroin with cyanogen bromide results in replacement of N—CH$_3$ by N—CN, and hydrolysis of the product yields the secondary base normorphine, from which a variety of N-substituted bases may be produced. The activities of these bases in the usual pharmacological tests decreases initially as the size of the group attached to nitrogen increases, but then rises again, reaching a peak at about C$_6$. The most active compound in this series is N-β-phenylethylnormorphine, which is about 14 times as potent as the original alkaloid and is the most active of all the simple morphine derivatives. With a three- or four-carbon unsaturated chain attached to the nitrogen, bases are obtained having an action antagonistic to that of morphine, although some of these are effective analgesics in man. This important group is discussed further below.

Simpler fragments of the morphine structure. Several groups of substances that may formally be regarded as being fragments of the morphine molecule have been synthesized, and the factors governing high morphine-like activity are now well known. The structure-activity relationships in the groups may be summarized as follows.

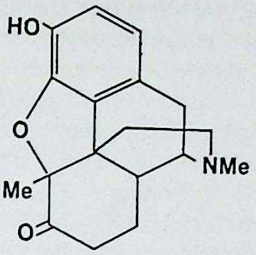
The morphinans [xviii–xxi] are bases representing the fundamental carbon skeleton of morphine without the 4,5-oxygen bridge. The basic compound, N-methylmorphinan, is a weak analgesic; and optimum activity is found in the 3-hydroxy-series, laevorphanol (iv) being about twice as potent as morphine. Ring C is *cis*-fused to ring B, and the corresponding *trans* isomers are more active [22]. B/C *trans* morphine has not yet been prepared, but the recent work of M. Gates and M. S. Shepard [23] has opened up a possible route to it. The results of varying the substituent on the nitrogen atom in this series have been found to be similar to that in the morphine group, and the most active compound is N-β-furylethyl-3-hydroxymorphinan (v), which is about 200 times as potent as morphine [21].


HO
2
3
1
A
4
11
12
10
O
B
NMe
13
5
14
9
C
6
8
HO
7

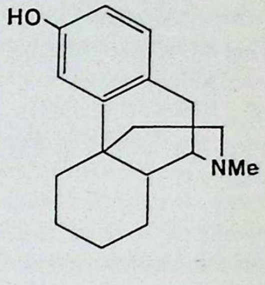
I


MeO
O
NMe
OAc
O

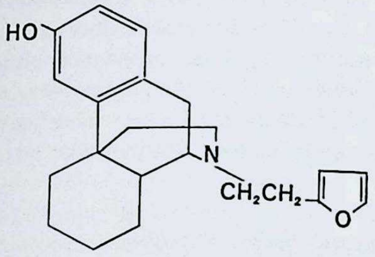
II



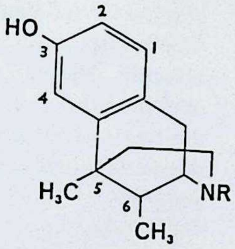
III



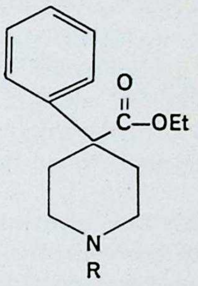
IV


HO
N
CH₂CH₂
O

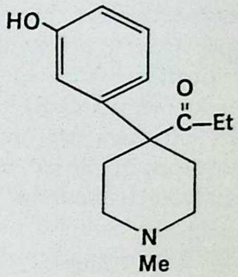
V


HO
2
3
1
4
H₃C
5
6
NR
CH₃

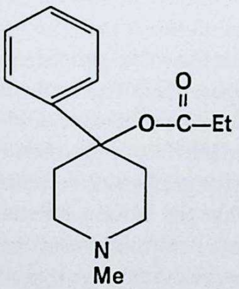
VI



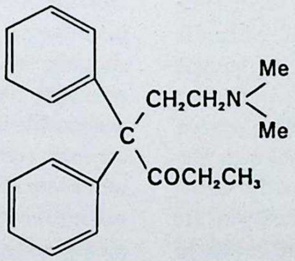
VII



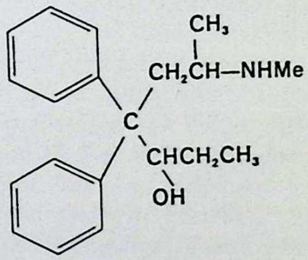
VIII



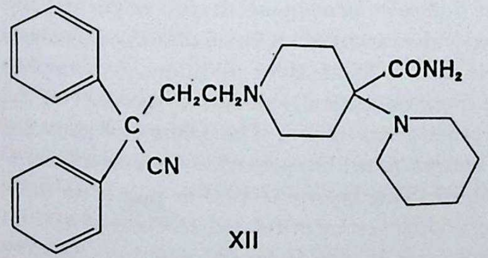
IX


Me
CH₂CH₂N
Me
C
COCH₂CH₃

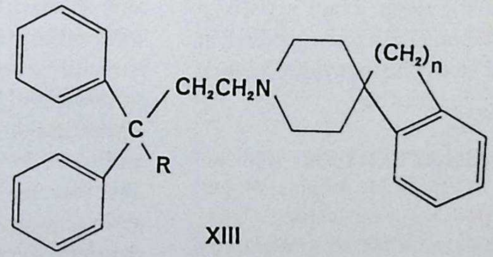
X


CH₃
CH₂CH—NHMe
C
CHCH₂CH₃
OH

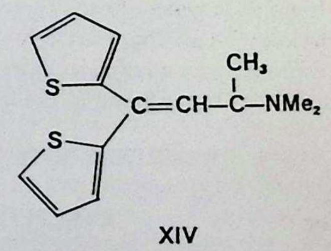
XI


CH₂CH₂N
CONH₂
C
N
CN

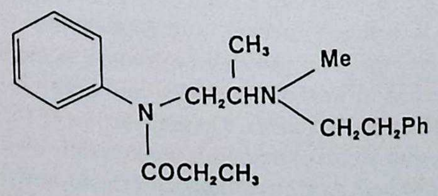
XII


(CH₂)ₙ
CH₂CH₂N
C
R

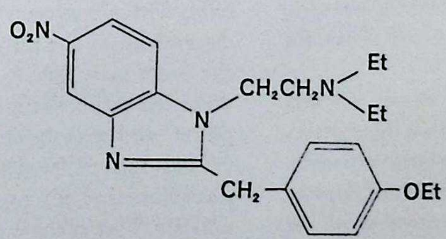
XIII


CH₃
S
C=CH—C—NMe₂
S

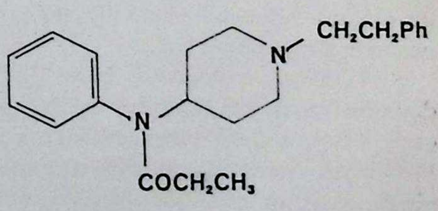
XIV


CH₃
Me
N
CH₂CHN
CH₂CH₂Ph
COCH₂CH₃

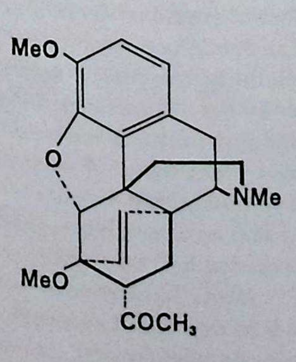
XV


O₂N
Et
CH₂CH₂N
Et
N
N
CH₂
OEt

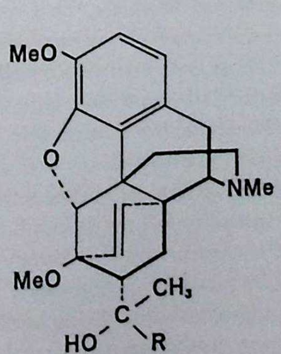
XVI


CH₂CH₂Ph
N
N
COCH₂CH₃

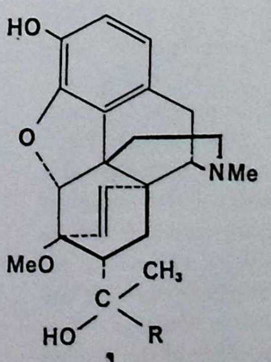
XVII


MeO
O
NMe
MeO
COCH₃

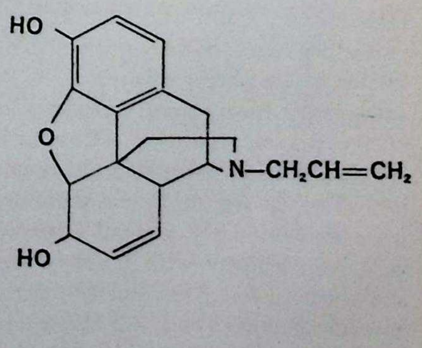
XVIII


MeO
O
NMe
MeO
CH₃
C
HO
R

XIX


HO
O
NMe
MeO
CH₃
C
HO
R

XX


HO
O
N—CH₂CH=CH₂
HO

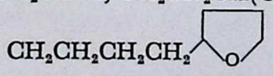
XXI

The benzomorphans [4, 5, 24, 25] show that the alicyclic ring C of the morphinan system is not necessary for high analgesic potency, because the simpler benzomorphans (VI) also possess this property. Some residue of ring C attached to positions 6 and 7 is, however, essential for maximum activity, as also is a phenolic hydroxyl at position 3. As in the morphinan series, activity is highest with the side-chain at C-5 and the hydrogen atom at C-6 in the *trans* relationship. Variation of the nature of the substituent on the nitrogen atom has the same effect as that already described in the morphine and morphinan series. With the 5,6-*cis* compounds the N-methyl compound metazocine (VI, R=Me) is as potent as morphine, and phenazocine (VI, $R=CH_2CH_2Ph$) is twenty times as potent. After extensive tests in monkeys, the last-named base was originally claimed to be appreciably less addictive and to produce substantially less respiratory depression than morphine [4], but these encouraging observations have not been substantiated in clinical use.

4-Phenylpiperidines [26–29] result from removal of the methylene group at position 8 of the benzomorphan skeleton (VI). A large number of analgesics have been prepared, some of very high potency, having this fundamental skeleton. In this series the carbon atom at position 4 of the piperidine bears no hydrogen in the active compounds, but, contrary to experience in the morphinan and benzomorphan series, a phenolic hydroxyl is not essential for even the highest activity. The first representative of this group to be produced was pethidine (VII, R=Me), which was originally prepared during a search for an atropine-like drug [30]. It has about one-seventh the potency of morphine and has become very widely used, especially in obstetrics, in which it owes its popularity to its alleged antispasmodic action; for this, however, the evidence is tenuous and far from convincing. Early work indicated that variation of the group attached to nitrogen led to loss of activity but this is generally the reverse of the truth. High activity is found with an approximately six-carbon chain or its equivalent containing an ether group attached to the nitrogen, the highest activities being found in the bases (VII) having

$$R=CH_2CH_2COPh,\ CH_2CH_2CH(OH)Ph,$$

or



$CH_2CH_2CH_2CH_2$

[28, 29]. High potency is also found in the related 4-ketones, for example ketobemidone (VIII), and in the reversed esters, for example the prodines (IX).

3,3-Diphenylpropylamine bases [5], like those already described, contain a phenyl group linked through a quaternary carbon atom and two other carbons to a tertiary basic nitrogen. For highest activity, the quaternary atom bears an ester, amide, nitrile, keto, or alcoholic substituent. The prototype of the series is methadone (X), which is one of the few commercially available drugs to which addiction may develop by oral administration. Until recently, it was only in this series that secondary bases, namely normethadol (XI) and its esters, were known to be potent analgesics. P. A. J. Janssen and his coworkers [6] have recently produced some interesting variants of the parent structure. These include pirinitramide (XII), which has also some structural resemblance to pethidine and is said to be less addictive than morphine; and a series of spiro-compounds of general structure (XIII) which combine very high activity with a very long duration of action. An analogous series of compounds, the thiambutenes having the general structure (XIV), are also morphine-like in action. They contain the isosteric thiophene nucleus in place of the phenyl groups and a quasi-quaternary carbon atom linked to its neighbour by a double bond [31].

2-Anilinoethylamines [4, 6] result from replacement of the quaternary carbon atom of the methadone (X) structure by a tertiary non-basic nitrogen atom. This leads to another group of analgesics of which the base diampromide (XV) is about as potent as morphine. Skeletally related to this series is the group of very highly potent benziminazoles, of which the most active base is etonitazine (XVI), one thousand times as potent as morphine [32]. Unfortunately, these compounds cause severe respiratory depression and have found no clinical use. It has also been found that analgesic activity is retained if the distance between the two nitrogen atoms is increased; Janssen has reported that base R4263 (XVII) is about 2000–5000 times as potent as morphine. Compounds of this type represent the first major analgesics in which there are more than three atoms between the aromatic nucleus and the basic nitrogen [6].

Compounds having molecules more complex than morphine [33–36]

None of the compounds of the various series described above has proved to be a notable improvement on morphine, and no worth-while separation of desired and undesired effects has been found. The writer felt that this was because all represent simpler, and therefore more flexible, molecules that would probably be at least as acceptable as, and possibly more acceptable than, morphine itself at the presumably closely similar receptors associated with analgesia, respiratory depression, and so on. This assumes that the various effects of morphine do not represent one inseparable complex resulting from action at a single site, which is an equally tenable hypothesis. It seemed reasonable, therefore, to make compounds more complex than morphine, which, by virtue of their greater complexity, rigidity, and substantially different peripheral shape, might be unacceptable at some of the receptors. A series of such compounds is readily available by the Diels-Alder addition of dienophils to thebaine [3], a pharmaceutically useless by-product of the extraction of morphine from opium. The adduct of thebaine and methyl vinyl ketone (XVIII) has proved to be a most favourable starting material. This ketone is itself as potent an analgesic as morphine, and the series of related alcohols of general structure (XIX), prepared by the Grignard reaction, provides bases with a very wide range of potency. The activity is highly dependent on the nature of the group R, the most active being the bases in which R is *n*-propyl, benzyl, and β-phenylethyl; these are respectively 95, 150, and 550 times as potent as morphine, and greatly exceed in analgesic activity anything previously derived from the alkaloid series. Demethylation of the phenol methyl ether group yields compounds of extremely high activity. Two of these, the base (XX, R=*i*-Am) and the O-acetyl ester of the base (XX, R=*n*-Pr), are the most potent analgesics so far reported, being approximately 10 000 times as active as morphine. In this series a number of secondary bases also have been shown to have very high analgesic potency.

These compounds are all relatively non-toxic, their therapeutic indexes being generally in the region 25 000–50 000. Effectively, the therapeutic index is the ratio of the lethal dose to the effective dose; it is a convenient index of drug safety. Since deaths from narcotic overdosage are usually due to respiratory failure, it was hoped that an appreciable separation of analgesia and respiratory depression had been achieved, but unfortunately experiments on humans have shown that the compound most closely studied is as depressant as morphine at doses producing an equal degree of analgesia [37]. As with all the previously described analgesics, the bases of this series are all highly addictive, and it is clear that the usual pharmacological screening tests simply ensure the selection of typically morphine-like compounds. The results further suggest that the pharmacological properties of morphine are in fact limited as an inseparable complex, and that the search for the ideal analgesic must be made along other lines.

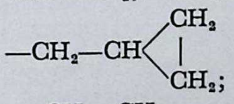
Morphine antagonists as clinical analgesics

As previously mentioned, N-allylnormorphine (nalorphine) (XXI) has long been known to antagonize all the effects of morphine. For this reason it is a valuable antidote in cases of narcotic overdosage, when the effects are dramatically rapid. Nothing is known with certainty about the mode of action of antagonists; possibly they successfully compete with morphine and similar drugs for occupation of the receptor surface. The replacement of N-methyl in the morphinans, benzomorphans, and bases of the series represented by (XIX) and (XX) also leads to morphine antagonists, some of which are more potent than nalorphine itself. These antagonists precipitate acute withdrawal symptoms when administered to narcotic addicts, and have not been shown to produce primary addiction themselves when administered over a prolonged period to animals and humans. They are inactive as analgesics when tested in animals by the tail-pressure or hot-plate method, and it was a surprising accidental finding that nalorphine is as potent an analgesic as morphine against post-operative pain in humans [38]. Unfortunately, the drug produces bizarre and often distressing mental effects (these include hallucinations), and it is accordingly of little clinical value as an analgesic. However, since it is not an addictive drug, and since it reverses respiratory depression caused by morphine and causes no untoward depression when administered alone, an intensive search is being made among similar compounds in the hope of finding one that does not cause the curious mental effects. This is probably the most important lead so far obtained in the search for the ideal analgesic.

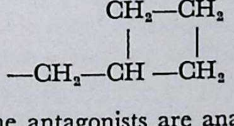
The range of substituents on the nitrogen atom resulting in non-addictive nalorphine-like compounds is very narrow. So far, it appears to be confined to unsaturated three- and four-carbon chains, and the following groups seem to be the most effective:

$$—CH_2CH{=}CH_2;\ —CH_2—CMe{=}CH_2;\ —CH_2CH{=}CMe_2;$$

$—CH_2C{\equiv}CH;$



and



Not all the morphine antagonists are analgesics, however; laevallorphan (the N-allyl-analogue of laevorphanol (IV)), for example, appears to be inactive in this respect. The pharmacological selection of the compounds, which are inactive by the usual tests, presents a problem, therefore, though some success has been achieved by selecting compounds inactive in such tests but active in the 'anti-writhing test' previously mentioned [12]. The number of starting materials in the morphine, morphinan, and benzomorphan series is limited, but a wide range of compounds of types (XIX) and (XX) has been, and is being, studied. It is reasonable to hope that out of work along these lines a safe, non-addictive analgesic will eventually emerge.

References

[1] Sertürner, F. A. W. *Trommsdorff's Journal der Pharmazie*, **13**, 234, 1805.

[2] Colman Green, G. *Nature, Lond.*, **157**, 465, 1946.

[3] Braenden, O., Eddy, N. B., Halbach, H., and Wolff, P. O. *Bull. World Hlth Org.*, **2**, 193, 1949; **10**, 1003, 1954; **13**, 937, 1955; **14**, 353, 1956.

[4] Eddy, N. B. *Chem. and Ind.*, 1462, 1949.

[5] Janssen, P. A. J. 'Synthetic Analgesics', Part 1. Pergamon Press, Oxford. 1960.

[6] *Idem. Brit. J. Anaesth.*, **34**, 260, 1962.

[7] Janssen, P. A. J. and Eddy, N. B. *J. med. pharm. Chem.*, **2**, 31, 1960.

[8] Beecher, H. K. *Pharm. Rev.*, **9**, 59, 1957.

[9] Siegmund, E., Cadmus, R., and Lu, G. *Proc. Soc. exp. Biol., N.Y.*, **95**, 729, 1957.

[10] Hendershot, L. C. and Forsaith, J. *J. Pharmacol.*, **125**, 237, 1959.

[11] Lister, R. E. (Unpublished work.)

[12] *Idem. Nature, Lond.* (In the press.)

[13] Beckett, A. H. *Progr. Drug Res.*, **1**, 527, 1959.

[14] Beckett, A. H., Casy, A. F., and Harper, N. J. *J. Pharm. Pharmacol.*, **8**, 874, 1956.

[15] Rapoport, H., Elison, C., Elliott, H. W., and Look, M. *J. med. pharm. Chem.*, **6**, 237, 1963.

[16] Orahovats, P. D. and Lehman, E. G. *J. Pharmacol.*, **122**, 58A, 1958.

[17] Fraser, H. *Ibid.*, **122**, 359, 1958.

[18] Grewe, R., Mondon, A., and Nolte, E. *Liebigs Ann.*, **564**, 161, 1949.

[19] Gates, M. and Tschudi, G. *J. Amer. chem. Soc.*, **78**, 1380, 1956.

[20] Schnider, O. and Grussner, A. *Helv. chim. Acta*, **32**, 829, 1949; **34**, 2211, 1951.

[21] Eddy, N. B., Besendorf, H., and Pellmont, B. *Bull. Narcotics*, **10**, No. 4, 23, 1958.

[22] Gates, M., Tschudi, G., and Webb, W. G. *J. Amer. chem. Soc.*, **80**, 1186, 1958.

[23] Gates, M. and Shepard, M. S. *Ibid.*, **84**, 4125, 1962.

[24] May, E. L. *et al. J. org. Chem.*, **22**, 1366, 1957; **24**, 1432, 1435, 1959; **25**, 984, 1960.

[25] May, E. L. *et al. Ibid.*, **27**, 245, 2144, 2554, 1962.

[26] Beckett, A. H. and Casy, A. F. *Bull. Narcotics*, **9**, No. 4, 37, 1957.

[27] Beckett, A. H., Casy, A. F., and Harper, N. J. *Chem. and Ind.*, **19**, 1959.

[28] Janssen, P. A. J. *et al. J. med. pharm. Chem.*, **1**, 105, 281, 299, 301, 1959; **2**, 271, 1960.

[29] Stern, E. S. *et al. J. chem. Soc.*, 4088, 1956; 3062, 3065, 1958; 2103, 1960.

[30] Eisleb, O. and Schaumann, O. *Dtsch. med. Wschr.*, **65**, 967, 1939; *Arch. exp. path. Pharm.*, **196**, 109, 1940.

[31] Wright, W. B., Brabander, H. J., and Hardy, R. A. *J. Amer. chem. Soc.*, **81**, 1518, 1959.

[32] Hunger, A., Kebrle, J., Rossi, A., and Hofmann, K. *Helv. chim. Acta*, **43**, 1032, 1960.

[33] Bentley, K. W. and Thomas, A. F. *J. chem. Soc.*, 1863, 1956.

[34] Bentley, K. W. and Hardy, D. G. *Proc. chem. Soc.*, 220, 1963.

[35] British Patents Nos. 902,659, 925,723, 937,214.

[36] British Patent Applications 8219 : 62 and 42668 : 62.

[37] Campbell, D., Lister, R. E., and McNicol, G. *J. clin. Pharmacol.* (In the press.)

[38] Winter, C., Orahovats, D., and Lehman, E. G. *Arch. int. Pharmacodyn.*, **110**, 186, 1957.

Glaciology

G. de Q. Robin

The water solidified in polar ice sheets is extremely pure, and this means that it is reasonable to attempt to compare laboratory studies of the properties of ice with field studies. This article describes the development of the study of ice as a plastic material, and the attempts to confirm theories by investigating present-day glaciers and the traces of prehistoric glaciers. The flow studies also provide useful experience for investigating any flow that might occur in the Earth's crust.

Glaciology deals with the study of ice in its many forms. These forms include the small ice crystals in the high atmosphere, the ice sheets that cover Antarctica and Greenland, the pack ice that covers polar oceans, river ice, lake ice, and the ice in permanently frozen ground. Extensions of the subject deal with such topics as erosion and deposition of material by ice, and the rate of colonization by plants of areas from which ice has receded. However, this article will deal mainly with the central problem of physical glaciology, with the flow of glaciers, and with some applications of glaciological studies to other disciplines.

Ice, as deposited in nature, is very pure. In the Antarctic ice sheet, for example, solid material trapped or dissolved in the ice has been shown by recent studies to constitute only a few parts per million of the total mass. Thus we should be able to study the mechanical and flow properties of ice in detail in the laboratory, and apply these results to field studies. For example, the variation with temperature of the velocity of compressional (P) elastic waves in ice, as measured in the laboratory by the author [1], shows close agreement with extensive field observations made in Antarctica between 1956 and 1963 (figure 1). Both sets of results apply to ice in which crystals have a random orientation, such as exist in the upper portions of thick ice sheets. However, ice cores obtained closer to the bottom of a glacier tend to show strong crystal orientations (figure 2), and we must view with caution the extrapolation throughout the whole thickness of a glacier of results for polycrystalline ice.

The flow of ice

Although many earlier but inconclusive experiments had been carried out, the first extensive and carefully controlled measurements of the flow of ice were made in Cambridge, by measuring the rate of deformation of cylinders of polycrystalline ice under uniaxial compression (figure 3, curve 1) [2]. Over the stress range 1–10 kg cm^{-2}, the relationship between stress and strain appears to be described adequately by the equation

$$\dot{\varepsilon}=k\sigma^n \qquad \ldots.(1)$$

where $\dot{\varepsilon}$ is the strain rate, σ the stress, and n and k are constants.

The value of the exponent n was found, after eliminating a transient term, to be 4·2 for the minimum creep rate. Other types of experiment have been carried out to check the value of this exponent, particularly by measuring the rate of closure of cylindrical tunnels and boreholes in glaciers, and in general, n has been found to lie within the range 2·8–4·2. However, studies of single crystals by S. Steineman [3] indicate that for ice deformed along the plane of easiest glide, n may fall to around 1·5 for low stresses. In a glacier, single crystals obviously cannot be deformed indefinitely. There is a continuous growth and decay of individual crystals, with a strong tendency for crystals in the lowest part of a glacier to grow with glide planes parallel to the plane of shear stress. This is shown by figure 2. Another significant experiment on ice flow shows [4] that, for a single crystal, the rate of deformation depends only on the shear stress, and is independent of the hydrostatic pressure, provided the difference between the ice temperature and the melting point remains constant. Thus ice at −1·0°C under atmospheric pressure deforms at the same rate as, say, ice at −2·8°C under 300 atmospheres pressure when subject to the same shear stress.

Applications

Application of the laws of ice flow to the study of glaciers can be divided into two parts. These are the internal deformation of the ice of the glacier, and the processes governing the sliding of a glacier over its bed. The first process will be discussed in more detail, since it is always of importance; the latter vary greatly in magnitude according to location. Until 1950, theoretical studies of the flow of glaciers had treated ice as a liquid with a high, but constant, Newtonian viscosity (figure 3, curve 2). In 1951, J. F. Nye attacked the theoretical problem by treating ice as an ideal plastic (figure 3, curve 3) which deforms indefinitely once the stress exceeds a certain value, τ_0 [5]. This assumption produced a useful simplification in

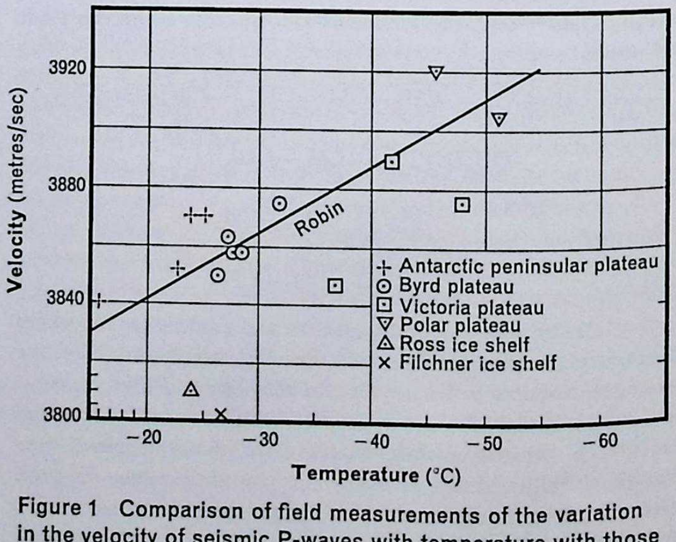


Figure 1 Comparison of field measurements of the variation in the velocity of seismic P-waves with temperature with those previously obtained from the laboratory measurements of Robin. The temperatures plotted are those at about 10 m depth. On inland ice, 10 m temperatures should be close to those at the depth for which velocities are given (100–400 m), but the apparently low velocity of the two results on ice shelves is probably due to a more rapid rise of temperature with depth than on inland ice. Note the open velocity scale used [20]. From C. R. Bentley (in press).

G. de Q. Robin, *M.A., Ph.D.*

Was born in 1921 in Melbourne, Australia, where he studied physics at the University. After wartime naval service, he worked in the University of Birmingham, in Antarctica, and at the Australian National University before his appointment as Director of the Scott Polar Research Institute in 1958.

His research has included studies of thickness and temperature distribution in large ice sheets, and ocean wave penetration into floating pack ice.

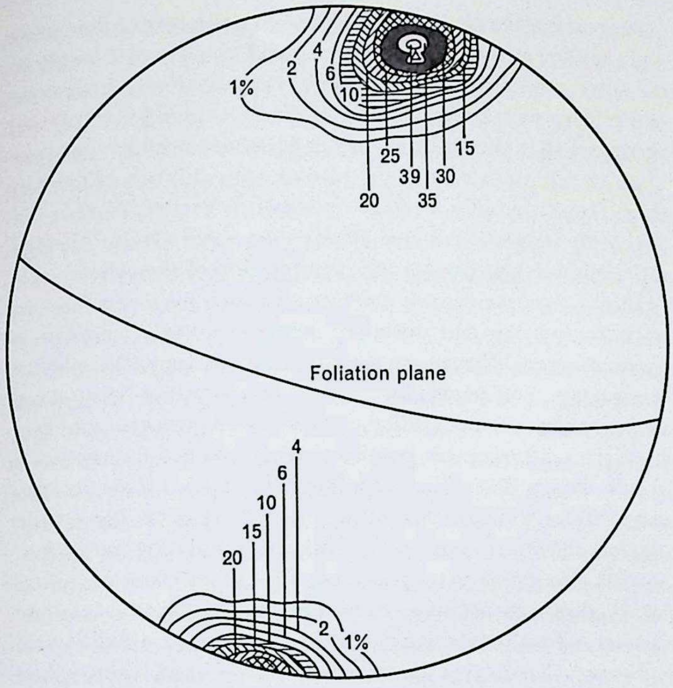


Figure 2 Fabric diagram of an ice section from near the bottom of a polar ice sheet (Tuto ice tunnel, north-west Greenland). The orientation of ice optic axes is plotted on the lower hemisphere of a Schmidt equal-area projection. Contours show the number of optic axes lying within 1 per cent of the area; the small triangle shows the position of the normal to the foliation plane, the plane of visible banding structure in the ice along which shearing has apparently taken place. Up to 39 per cent of the optic axes are concentrated in 1 per cent of the area almost normal to the foliation plane [24].

the analysis. Thus in figure 4, if we assume the simplified condition of laminar flow, where all shear is parallel to the top and bottom surfaces of the glacier, we find the shear stress on the bed, which must be equal to τ_0 for flow to occur, is given by

$$\tau_0 = \rho g h \sin \alpha \qquad \ldots.(2)$$

Thus $h = h_0 \sin \alpha$, where $h_0 = \rho g/\tau_0$.

In other words, under these conditions the ice thickness is inversely proportional to the surface slope α (taking $\alpha = \sin \alpha$ for small angles). A refinement [6] derived expressions for glacier thickness and flow, assuming the simplified flow model in figure 4 but using a realistic flow law. However, the most significant advance was that made by assuming the simplified flow law $\tau = \tau_0$ at all points in a glacier (figure 3, curve 3), but without making any major assumptions about the way in which ice deformed within the glacier. Later work [7] analysed the problem, using both a realistic flow law and imposing relatively few restrictions on the mode of ice deformation.

The essential points of glacier flow are perhaps explained most easily by assuming that the flow is plastic ($\tau = \tau_0$). Equations for flow can be derived from L. Prandtl's solution [8] for the flow of a plastic material compressed between plane parallel plates with rough surfaces, so that no sliding occurs against the face of the plate. The solution for glacier flow is obtained by adding a gravitational potential proportional to the position in a glacier. The surface velocity of the glacier is that of the median plane between the plates, and the full velocity solution is:

$$u = C \pm V \cdot \frac{x}{h} + 2V\sqrt{\left\{1 - \left(1 - \frac{y}{h}\right)^2\right\}} \qquad \ldots.(3)$$

$$v = \mp V \cdot \frac{y}{h} \qquad \ldots.(4)$$

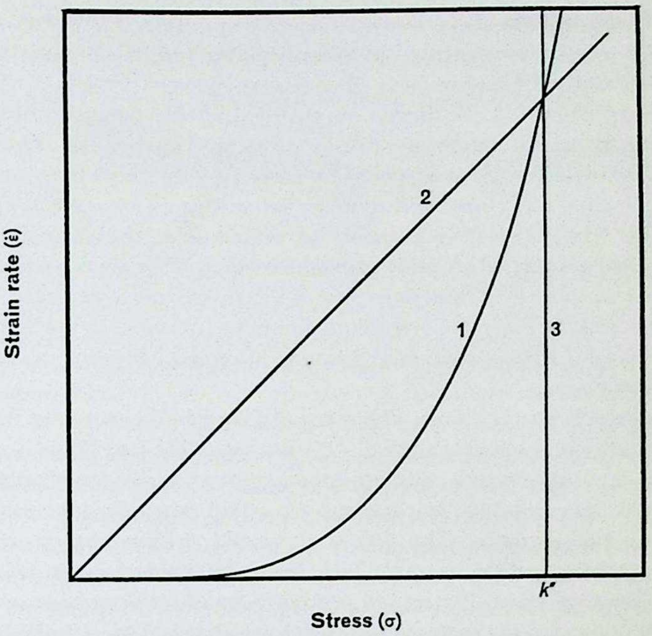


Figure 3 Types of flow law for ice used in theoretical studies of the flow of glaciers. (1) Glen [2]: $\dot{\varepsilon} = k\sigma^n$, where $n \approx 4{\cdot}2$ from laboratory determinations. (2) Newtonian viscosity: $\dot{\varepsilon} = k'\sigma$. (3) Plastic: $\sigma = k''$, used by Nye [5]. k, k', and k'' are constants.

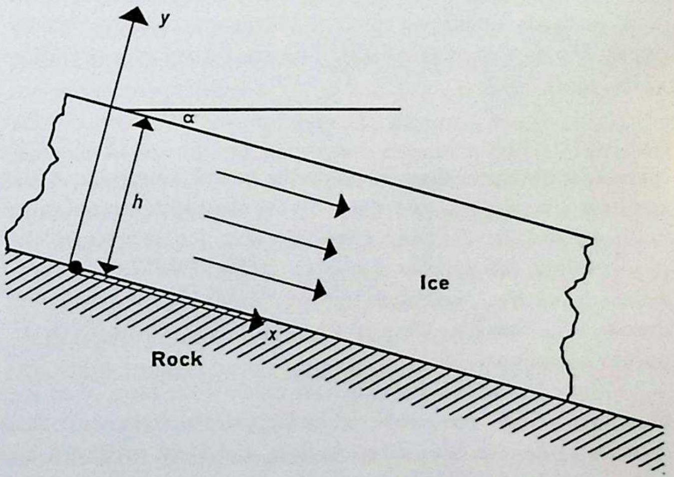


Figure 4 Laminar flow. All particles of ice flow in planes parallel to that of the glacier bed. α is the angle the rock bed makes with the horizontal.

where u and v are the velocity components in the x and y directions (as in figure 4), h is the thickness of the glacier, and C and V are constants.

It is readily seen that, with a plastic material between plates, solutions to the problem can be obtained only if there is longitudinal compression or extension taking place to permit conservation of volume as the plates are forced together or pushed apart. If one applies the plastic model to a glacier, there must also be longitudinal extension or compression throughout the glacier unless the precise conditions of laminar flow are met. The two types of flow were termed extensive and compressive flow, respectively, by Nye. The plastic glacier thus deforms by shearing parallel to the bed in the lowest layers. Close to the surface, the difference between longitudinal and vertical stresses results in shear planes meeting the surface at an angle of 45 degrees. This simplified picture of the mechanism has been shown [7] to be approximately correct. Some detailed improvements to Nye's work have been made since that date, and field evidence shows that this picture applies generally.

Confirmation

The most extensive attempt to confirm Nye's conclusions on the velocity distribution in a glacier was reported by W. S. B. Paterson and J. C. Savage in 1963 [9]. They obtained systematic measurements of velocity and strain rates in the Athabasca Glacier, Canada, firstly by measuring surface movements of a system of stakes and secondly by measuring movements at depth by means of eight boreholes, mainly along the centre line of the glacier. The simple geometrical form of the glacier made it possible to fit the flow law of ice to the movement. The underlying theory of Nye was found to be substantially correct, and the stress-strain rate relation fitted an exponent, *n*, of about 4·2.

Much less is known about the sliding of a glacier over its bed than about the internal deformation of the ice. There is a lack of sufficient laboratory experiments to verify hypotheses with any certainty, but it seems clear that two main processes are taking place. The first is the rapid deformation of ice around any obstacles on the bed, due to the rapid concentration of local stresses. The second process is the effect of pressure on melting and of regelation, in which ice that is subject to high pressure on the upstream side of an obstacle melts because of the change of melting point with pressure. The ice is thought to refreeze on the downstream side of the obstacle, the heat liberated during refreezing being conducted through the obstacle to the upstream face so as to give a continuous process. J. Weertman developed [10] a rough quantitative theory depending on these two effects. This gave a velocity of sliding of ice, *u*, of

$$u = (\tau/A)^m \qquad \ldots.(5)$$

where τ is the mean shear stress on the bed of the glacier, A is a constant, and $m = \frac{1}{2}(n+1)$ where n is the exponent in equation 1.

Nye explained [11] the form of polar ice sheets on the hypothesis of this process of bottom sliding. The derived form reduced to that predicted by his 'plastic' theory when m approached infinity. This is the point that corresponds to plastic behaviour.

A simple test of this hypothesis is to see if the form of an ice sheet agrees with that predicted by hypothesis. In general, this appears to be the case. For example, the form predicted by Nye [11] for a two-dimensional polar ice sheet on the basis of sliding is

$$\left(\frac{h}{H}\right)^{2\,4} + \left(\frac{x}{L}\right)^{14} = 1 \qquad \ldots.(6)$$

where H is the height, in kilometres, of a sheet of width $2L$ km, and h is the height at a distance x km from the centre.

B. Imbert, in reporting [12] the form of a section of the Antarctic ice sheet in Adelie Land, fitted his observations to the formula for the ellipse

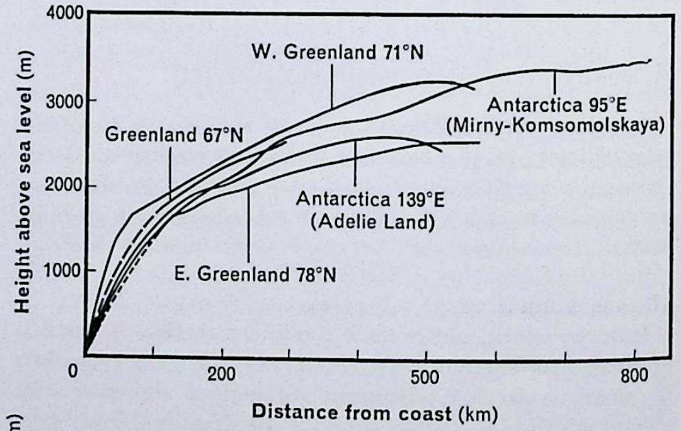
$$\left(\frac{h}{2{\cdot}68}\right)^2 + \left(\frac{x}{550}\right)^2 = 1 \qquad \ldots.(7)$$

Although the two formulae differ, they both predict surface heights that agree with the observed heights to within 100 m. The surface form of practically all glaciers and ice sheets appears readily explicable in terms of current hypotheses if the shear stresses range from 0·3 to 2·0 bars. It is not, however, possible to fit the predicted surface form precisely to the theory, partly because of the great number of data necessary to plot a glacier bed in detail, and also because the theory is not yet sufficiently precise in itself. For example, variations of longitudinal stress along a glacier are not yet incorporated into the theory.

A consequence of the rapid increase in the rate of flow of ice with applied stress is that the profiles of glaciers and ice sheets on approximately horizontal beds vary relatively little from one glacier to another. In two of Nye's papers [5, 11] it was predicted that the thickness of an ice sheet would be proportional to the sixth root of the rate of accumulation. Therefore there should be only a relatively small difference, of about 16 per cent, between two ice sheets that were similar in form although one had double the precipitation of the other.

In practice, we cannot produce two such ice sheets for comparison, but we can compare profiles along a number of traverses over different ice sheets. Unfortunately, the effect of topography, and particularly of coastal mountain belts, causes variations between the different profiles in figure 5(*a*), but even so, there is a noticeable regularity of profile in spite of rates of accumulation that differ by a factor of at least three in some cases. More detailed knowledge is available on the smaller glaciers shown in figure 5(*b*), where accumulation varies from around 100 g/cm² per annum on the higher parts of Vatnajokull, down to negative values on the profile of Malaspina Glacier, where the section shown is wholly in the ablation area.

Enough has been said to make it clear that, provided we know the outline of an ice sheet and have an approximate idea of the underlying topography, we can make a good estimate of its surface form. This applies to present-day ice sheets and also to past ice sheets, since the basic physical processes involved do not change with time. The main difficulty in applying this knowledge to past ice sheets is that it is difficult to be certain



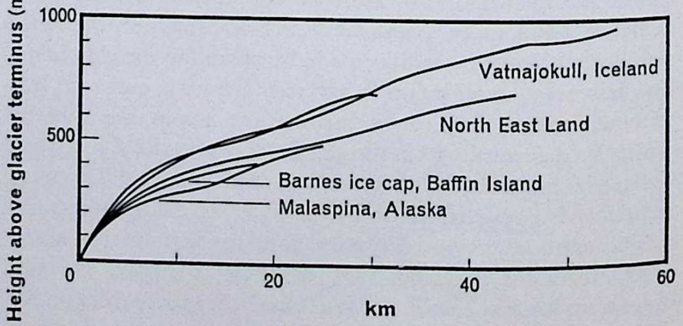


Figure 5(*a*) Surface profiles over sections of the continental ice sheets of Greenland and Antarctica, where the ice lies on rock generally within ±500 m of sea level. The Greenland profiles are obstructed by coastal mountain ranges of greater height, but the ice is discharged through deep valleys which cut through the mountains. Elevations shown are those above sea level.

Figure 5(*b*) Surface profiles of smaller ice sheets or ice caps over sections which have been shown to rest on an approximately horizontal bed. The heights shown are those above a horizontal plane starting from the ice–rock contact at the outer boundary of the ice cap.

of the extent of the ice at any one time. In general, the earlier the glaciation, the less certain we can be of its extent. We will therefore first consider the most recent ice sheet, the Fennoscandian ice sheet, which reached its maximum extent a few tens of thousands of years ago. Its outline, as shown in figure 6, is known with reasonable certainty from recent glacial deposits over Scandinavia. If we apply a typical profile from figure 5 to show the way in which the surface rose inland from the outer limits of the ice sheet, and make appropriate allowance for the fact that the mountains caused the ice surface to rise more rapidly on the western side, we find that the surface contours of the ice sheet must have been similar to those shown. The profile from around Trondheim to Helsinki would have been roughly as in figure 7. The extra weight of the ice should have depressed the Earth's crust by about one-third of the ice thickness, and the effect of this on the profile is shown.

In figure 7 it should be noted that the highest point of the ice sheet is displaced towards the west from the centre of the mass of the ice. This fits geological evidence, which shows that the centre of the outflow of the ice appeared to be displaced to the west of the centre of isostatic rebound. Another deduction from geological evidence comes from the maximum height of raised beaches, found to be about 250 m in central Sweden. This would imply that at least 750 m of ice formerly covered this area. However, the raised beach could not have been formed until the ice had disappeared, and since much rebound is already likely to have occurred as the ice melted, the maximum thickness of around 3000 m shown in figure 7 would not be inconsistent with the height of the raised beaches.

The evidence of the extent of the maximum Pleistocene glaciation over Europe is less certain. From what evidence there is, it seems that the ice must have reached a thickness of over 3000 m over the Baltic Republics of the U.S.S.R. One can apply estimates on this basis to other ice sheets such as those covering North America, Siberia, and other areas. One can then estimate the total volume of ice sheets resting on land at the time of maximum glaciation, and then, assuming that the maximum was reached at the same time in all areas, one can calculate the resultant lowering of mean sea level. Estimates of these quantities given by the author in a lecture in 1961 are compared in the following table with independent estimates published by W. L. Donn, W. R. Farrand, and M. Ewing [13]. The agreement is seen to be close, and moreover the latter authors report evidence of past sea levels which would fit these figures.

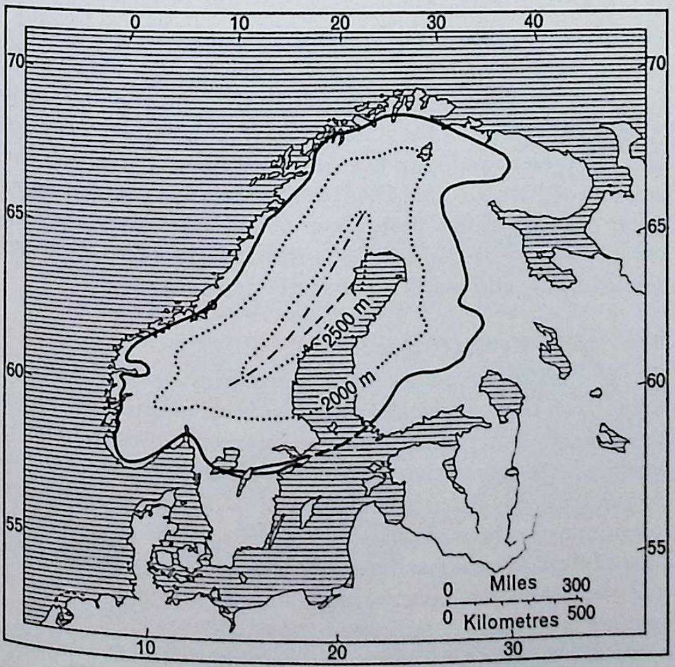


Figure 6 Approximate surface contours of the Fennoscandian ice sheet at the time of maximum glaciation in relation to the boundary of the ice sheet. Boundary according to Flint [21]. (— — — shows axis of area of greatest uplift after deglaciation. —·—·—· marks the most easterly known position of ice divide).

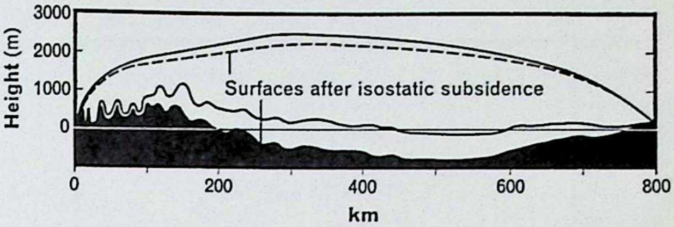


Figure 7 Estimated profile across the Fennoscandian ice sheet along a line running approximately from the present-day position of Trondheim to that of Helsinki. The profiles refer to the maximum glaciation at this time; the estimated ice and rock surfaces after allowing for isostatic subsidence are also shown.

Pleistocene ice volume and sea level lowering

Maximum glaciation ($km^3 \times 10^6$)	Robin (1961 lecture) (1)	Robin (1961 lecture) (2)	Donn, Farrand, and Ewing (1962) (1)	Donn, Farrand, and Ewing (1962) (2)
Total glacier ice	80	91	84	99
Existing glacier ice	28	28	25	35
Pleistocene ice excess	52	63	55	64
Water equivalent	47	57	50	58
Sea level lowering (metres)	129	157	137	159
Sea level change in metres after isostatic adjustment of ocean floor	83	105		

Other information from glaciers

In addition to providing general evidence of conditions during past ice ages, the existing ice sheets contain much detailed evidence of climate, precipitation, and other factors, during past centuries. This is preserved in the detailed stratigraphy of polar ice sheets. Only a small portion of this evidence has been recovered and evaluated so far, since the necessary techniques are difficult and time-consuming.

The simplest climatic record available from an ice sheet is that of the net annual accumulation. At suitable localities, there is a clear distinction between snow deposited during the summer and subject to slight intermittent melting and the drier and finer-grained snow of winter months, and with experience, glaciologists can deduce the thickness of past annual layers and hence determine the net precipitation. Thus at Byrd Station, Antarctica, variations of the net annual accumulation over the past centuries have been measured by direct inspection of ice cores down to depths of 90 m below the surface. Beyond this depth, the gradual compression and metamorphism of ice with depth made it impossible to follow the annual layering by a general inspection of the structure. However, the layering is not basically destroyed by this compression, and it may be identified by careful studies of the changes in the oxygen or hydrogen isotope ratio with depth provided that the samples are spaced at sufficiently close intervals (figure 8).

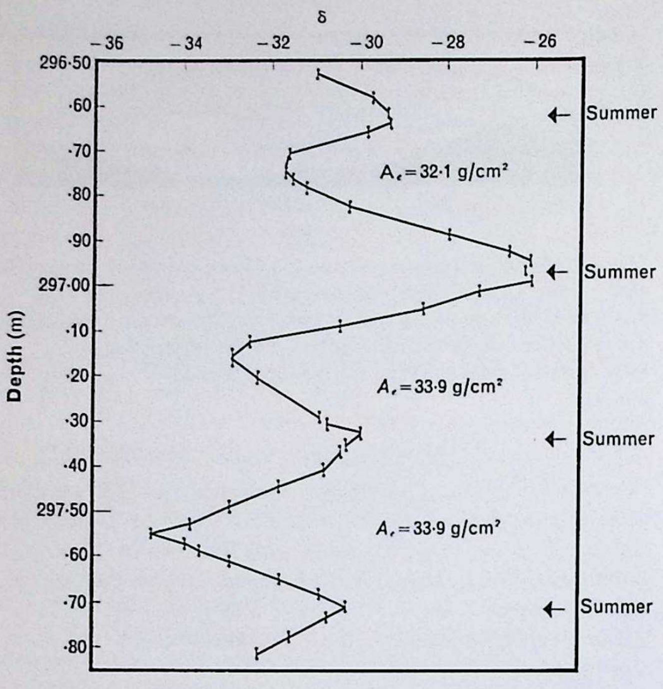


Figure 8 Annual layering in ice from a depth of 300 m in the ice sheet of north-west Greenland, shown by variations of the oxygen isotope ratio with depth [22, 23]. $\delta = [(R_s/R_o) - 1] \times 10^3$, where $R_s = {}^{18}O/{}^{16}O$ ratio of sample and $R_o = {}^{18}O/{}^{16}O$ ratio of standard (δ is zero for mean ocean water). A_e=net annual accumulation of ice [22].

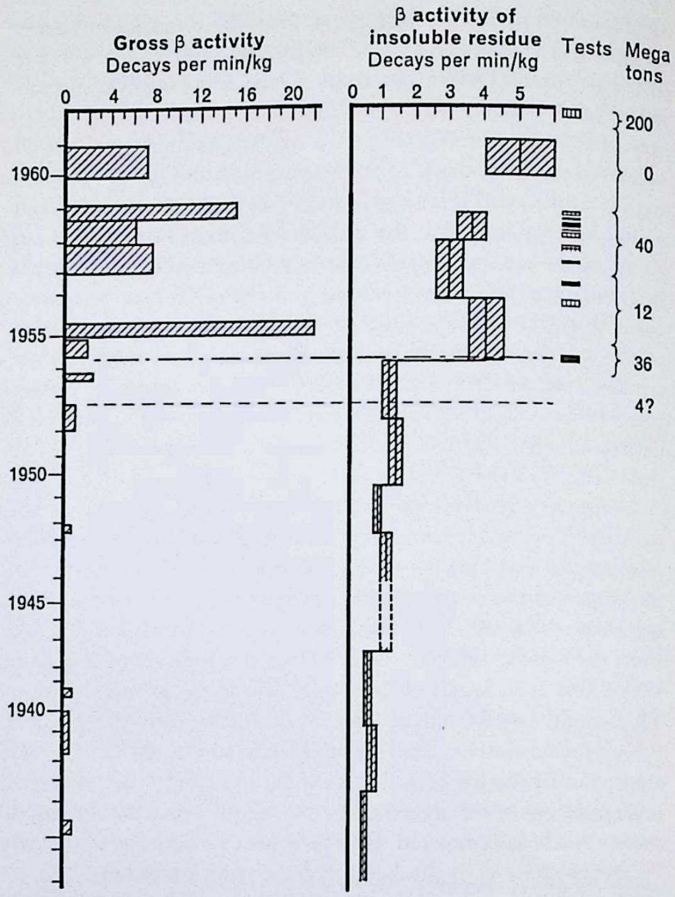


Figure 9 Radioactivity of firn (old snow) in Antarctica at Roi Baudouin station. (The central line of the right-hand set of values represents the β activity, the shaded areas to either side, the standard deviation.) The date of precipitation of the snow has been determined from stratigraphic data. On the right are shown the times of the main thermonuclear test series (Soviet tests in black, British and American tests shaded) and the estimated total fission yield from nuclear events in megatons [14].

Isotope studies are opening up new types of investigations of ice sheets and glaciers. Ratios of ^{18}O to ^{16}O and of ^2H to ^1H atoms in the ice depend on the temperature at which the precipitation was formed in the atmosphere. Thus the difference between summer and winter precipitation is marked permanently in many localities by the isotope ratios. However, in areas of low precipitation or of irregular deposition, annual layering is not always clear, and the system tends to break down. Alternative methods of dating the snow must therefore be sought. Possible approaches are the study of the variation with depth of naturally occurring radioactive isotopes, such as tritium and radon. However, a technique which appears more promising has been reported by E. Picciotto and S. Wilgain [14], and by W. W. Vickers [15]. They have shown that debris from thermonuclear explosions has contaminated the Antarctic ice sheet to such an extent that it is relatively easy to identify the firn (old snow) deposited in 1955, by its enhanced radioactivity (figure 9). If a reasonably rapid technique can be developed for identifying this layer in the field, the ambiguity of some precipitation measurements on the high Antarctic plateau will be eliminated.

Study of isotopic composition and of the minute amounts of impurities present in polar ice sheets can be of value in many fields other than glaciology. For example, the tritium content of ice may be studied to see if there is evidence of past changes of the amount of tritium in the atmosphere. During the past eight years, contamination from nuclear explosions has affected the tritium content of the atmosphere so much that it is necessary to study firn samples deposited before 1953 to search for any variations due to natural causes.

Polar ice sheets form suitable collecting grounds for examining the fall of cosmic and terrestrial dust. Minute magnetic spherules, which are believed to come from outside the Earth, are preserved in the ice, so that their rate of fall over long periods can be determined. One study [16] appears to indicate annual variations in their rate of deposit, a feature suggested by studies elsewhere. Great volcanic eruptions also leave a permanent horizon of volcanic ash in the ice: three such horizons have been identified in cores from the United States Little America V station on the Ross Ice Shelf [17]. Minute traces of salts are present and have been analysed quantitatively for the amounts of Na+, K+, and Cl+ [18]. The relative proportions of these elements differ from those of sea water, and a non-oceanic supply of Na ions is indicated, but more data are needed before the causes can be fully understood.

The 'Upper Mantle Project'

The preceding paragraphs set out some investigations in glaciology which bear specifically on other fields of investigation. A less specific but potentially very important contribution to geophysical studies arises from points made at the start of this paper. The slow deformation of ice in nature gives knowledge of the massive flow of solids on a scale intermediate between that of the laboratory and tectonic processes. In particular, the slow deformation of material in the mantle of the Earth must be similar to the flow of ice in many ways.

As the 'Upper Mantle Project', an international campaign for the study of the upper mantle, gets under way it is appropriate to draw the attention of geophysicists to recent advances in glaciological studies. Studies of the mantle are dealing with it in terms of 'convection currents' and the associated high values of 'viscosity'. Glaciological studies have now moved beyond such general terms. Instead of a 'Newtonian viscosity' for ice, we now use more precise laws when calculating the flow of ice sheets. Problems being studied in relation to glacier flow include the roles of crystal orientation, recrystallization under stress, and phase changes from water to ice. There are undoubtedly major differences between the behaviour of the material of the mantle and of ice, but much of the basic mechanics may well be similar. Thus Weertman modified [19] the basic equations he had formulated to describe the flow of floating ice shelves to fit the case of a spherical Earth. He showed how relatively small changes in properties of different parts of the mantle could generate the forces necessary to cause continental drift.

References

[1] Robin, G. de Q. 'Norwegian-British-Swedish Antarctic Expedition 1949–52', *Sci. Results*, **5**, Glac. III, Norsk Polarinstitutt.

[2] Glen, J. W. *Proc. roy. Soc.*, *A*, **228**, 519, 1955.

[3] Steineman, S. *J. Glaciol.*, **2**, 404, 1954.

[4] Rigsby, G. P. *Ibid.*, **3**, 273, 1958.

[5] Nye, J. F. *Proc. roy. Soc.*, *A.*, **207**, 554, 1951.

[6] *Idem. J. Glaciol.*, **2**, 82, 1952.

[7] *Idem. Proc. roy. Soc.*, *A.*, **239**, 113, 1957.

[8] Prandtl, L. *Z. angew. Math. Mech.*, **3**, 401, 1923.

[9] Paterson, W. S. B. and Savage, J. C. *J. geophys. Res.*, **68**, 4537, 1963.

[10] Weertman, J. *J. Glaciol.*, **3**, 33, 1957.

[11] Nye, J. F. *Ibid.*, **3**, 493, 1959.

[12] Imbert, B. *C.R. Acad. Sci.*, *Paris*, **248**, 576, 1959.

[13] Donn, W. L., Farrand, W. R., and Ewing, M. *J. Geol.*, **70**, 206, 1962.

[14] Picciotto, E. and Wilgain, S. *J. geophys. Res.*, **68**, 5965, 1963.

[15] Vickers, W. W. *Int. Ass. Sci. Hydrol. Publ. No.* 61, 199, 1963.

[16] Langway, C. C. *Ibid.*, *Publ. No.* 61, 189, 1963.

[17] Gow, A. J. *Ibid.*, *Publ. No.* 61, 272, 1963.

[18] Brocas, J. and Delwiche, R. *J. geophys. Res.*, **68**, 3999, 1963.

[19] Weertman, J. *Ibid.*, **67**, 1133, 1962.

[20] Bentley, C. R. *U.S. Acad. Sci.*, *IGY Symp.*, Los Angeles, 1963. (In press.)

[21] Flint, R. F. 'Glacial and Pleistocene Geology'. Wiley, New York. 1957.

[22] Langway, C. C. *Int. Ass. Sci. Hydrol. Publ. No.* 58, 101, 1962.

[23] Epstein, S. and Sharp, R. P. *Trans. Amer. Geophys. Un.*, **40**, 80, 1959.

[24] Rigsby, G. P. *J. Glaciol.*, **3**, 589, 1960.

Book reviews

PHILOSOPHY OF SCIENCE

Science and Sensibility, *by James R. Newman. Vol. I, pp. 372; Vol. II, pp. 309. George Allen and Unwin Ltd, London. 1963. 75s. net.*

The two volumes consist of forty essays, mainly long book reviews. These reviews are the occasion for the author to write about a number of themes of wider interest. They range from Greek science to causality in modern physics; and there are several articles discussing the effect of science on society. A large group of essays is concerned with biography; they span the period from Bacon to Einstein and Freud. Cardano, Clifford, and Ramanujan are some of the less well known personalities whose ideas are described by the author. The volumes contain many illustrations usually taken from the books under review.

Thus we have here a readable and urbanely presented survey of the basic ideas of modern science and of their impact on western thought. The essays give an example of the civilizing force that science has exerted on mankind and they prove that science not only makes sense but arouses our sensibilities as well. *Ernest H. Hutten*

COSMOLOGY

Interstellar Communication, a Collection of Reprints and Original Contributions, *edited by A. G. W. Cameron. Pp. xii+320. W. A. Benjamin Inc., New York. 1963. $9.35.*

The list of authors, ranging from the eminent Russian astrophysicist Shklovsky to American physicists and astronomers, including two Nobel prize winners, is a guarantee of the high scientific content of this book. The theme is that the solar system did not evolve as a result of a rare accident but is typical of the planetary systems that accompany most of the stars in the Universe. The fraction of these stars which have at least one planet in a stable condition for a sufficiently long term, a few thousand million years, to facilitate organic evolution is calculable, and the problem of the existence of life elsewhere becomes essentially a biological as distinct from an astronomical problem. The biological aspects are admirably treated by Melvin Calvin. The probability of the existence of many other communities in the cosmos leads to the problem of communication. Here the possibilities of optical lasers and radio-astronomy are fully explored. There is inevitably some repetition, since the book is largely a collection of articles published elsewhere; but this does not detract from the good impression created, and the presentation of the relevant background information in astronomy, radio-astronomy, and biology is admirable. *Bernard Lovell*

PHYSICAL DATA

Vapour Pressure of the Elements, *by An. N. Nesmeyanov (translated and edited by J. I. Carasso). Pp. vi+469. Infosearch Ltd, London; Cleaver-Hume Press Ltd, London. 1963. £5 5s. net.*

Vapour pressure measurements are rightly regarded as a source of important thermodynamic data. This excellent book, a translation of the Moscow edition of 1961, is more than a useful compilation of such data, for it contains a detailed account of experimental methods down to 10^{-10} mm of mercury and an adequate discussion of the theoretical background. Due attention is paid, for example, to the role of the evaporation coefficient and to polymerization in the vapour. The main section is a careful appraisal of the vapour pressure of the elements: hydrogen, oxygen, nitrogen and the 'inert' gases are not considered. The author points out that most of the data are tentative, the vapour pressure of even such a common element as mercury being in doubt. A much greater degree of uncertainty attaches to composition of the vapour of the elements. The last chapter is introductory to the tables at the end and gives theoretical and empirical methods for the integration of the Clausius-Clapeyron equation. In the tables, the constants of vapour pressure/temperature equations have been derived with the help of the best thermodynamic data by the use of the Strela computer at Moscow. It is easy to find one's way in this attractive book, but an index would have been an advantage. *R. S. Bradley*

COSMIC RAYS

Progress in Elementary Particle and Cosmic Ray Physics, *Vol. VI, edited by J. G. Wilson and S. A. Wouthuysen. Pp. x+344. North-Holland Publishing Co., Amsterdam. 1962. 80s. net.*

The editors decided at the time of the publication of the fifth volume of the series to separate the two groups of subjects covered—elementary particles and cosmic ray physics—into alternate volumes. They have carried out their intentions into effect with the volume under review. Volume VI is devoted exclusively to the study of cosmic-ray physics. But—and this shows the explosive development of the subject—the material planned to make up Volume VI has itself proved too extensive for inclusion in a single volume, and accordingly another volume—Volume VII—will shortly be issued, also covering cosmic rays.

The contents of Volume VI are; 'Cosmic-ray producing isotopes', by Lal and Peters; 'Time variations of low rigidity cosmic rays during the recent sunspot cycle', by W. R. Webber; and 'Geomagnetically trapped radiation', by Singer and Lenchek.

As usual, the standard of the contributions is of the highest. *A. Salam*

CHEMISTRY

Topics in Modern Chemistry, *by C. J. Mandleberg. Pp. xii+228. Cleaver-Hume Press Ltd, London. 1963. 37s. 6d. net.*

This is, as the author states in his preface, 'a very personal book' about chemistry. He devotes the first two of his eight chapters to generalities under the headings 'Models and methods' and 'Fact and fiction'. From these he jumps to an interesting and very useful account of fundamental particles, dealing in the main with nuclear structure. The next chapter, on chemical structure, lacks balance, both in the choice of content and in the level of treatment. That on solutions of electrolytes is largely about the chemistry of non-aqueous solutions and the behaviour of ligands. The treatment of the first of these topics is good; that of the second is superficial and unlikely to be of great value. The next two chapters on the Periodic Table and recent chemistry of some metals, and on some compounds of non-metals, contain a great deal of detailed descriptive chemistry, much of which has been published very recently. The final chapter, on some compounds of carbon says a little about each of a number of general topics, but the length precludes the satisfactory treatment of any of these. The book makes interesting general reading, especially because of the philosophical approach that is often adopted.

H. J. Emeléus

SURFACE-ACTIVE AGENTS

Colloidal Surfactants, some Physicochemical Properties, *by K. Shinoda, T. Nakagawa, B. Tamamushi, and T. Isemura. Pp. viii+310. Academic Press, New York and London. 1963. 92s. net.*

This co-operative monograph comprises four chapters on the physical chemistry of surface-active agents, under the headings 'The formation of micelles', 'Physicochemical studies in aqueous solutions of surface active agents', 'Adsorption', and 'Monomolecular layers'.

In their preface the authors mention that a considerable amount of research in this field has been carried out in Japan, but that it is not well known in other countries because many of the papers were published in Japanese, and even those written in English for the *Bulletin of the Chemical Society of Japan* are not widely read. The reason for this state of affairs may be guessed from a study of the present monograph. It is certainly a useful, clear, and generally accurate account of the academic world literature on synthetic surfactants. Several hundred relevant papers have been competently reviewed, and the text is written in better English than is often encountered in American textbooks. But what is sadly lacking is that precious spark of critical originality that might have thrown some fresh light on these tired old problems.

The book is useful for novices starting research on micellar colloids.

J. A. Kitchener

ANALYTICAL CHEMISTRY

Standard Methods of Chemical Analysis *(sixth edition), edited by Frank J. Welcher, Vols. IIA and IIB,* **Industrial and Natural Products and Noninstrumental Methods.** *Vol. IIA, pp. xiv+1282; Vol. IIB, pp. xi+1283–2613. D. Van Nostrand Co. Inc., Princeton, New Jersey and London. 1963. £9 9s. net.*

The new volume II of 'Scott', split into two parts for convenience of handling, covers much the same ground as the previous edition of 1939, with the exception that the small amount of instrumental material has been reserved for extended treatment in the forthcoming volume III.

The phenomenal expansion in, and development of, new materials and products in the intervening quarter-century is in some measure met by the introduction of ten new chapters which deal with subjects including pesticides, plastics, fertilizers, soils, and vitamins. However, the editor has been wisely restrictive in selecting new fields and has retained the flavour of previous editions. Single sources can no longer cope with the gamut of products and materials. This volume deals with the 'wet' or 'classical' chemical procedure for the analysis of technical materials, and takes some cognizance of the crescendo of activity of analytical-methods committees and standards institutions, with, of course, a strong bias towards American practice. The classical methods are of permanent importance but of dwindling appeal, and the presentation is still that of an art or skilled craft rather than that of a science. The extension of the section on general operations does little to alleviate this, and many of the illustrations, already of a classical cast in previous editions, are retained, although a great deal of the text has been rewritten, and maintains the standard of its predecessors.

E. Bishop

INORGANIC CHEMISTRY

Physical Inorganic Chemistry, *by M. J. Sienko and R. A. Plane. Pp. xi+166. W. A. Benjamin Inc., New York. 1963. $4.35 paper; $7.70 cloth.*

The study of modern inorganic chemistry requires a basic knowledge of physical chemistry, and the current approach is based firmly on quantum mechanics, thermodynamics, and kinetics. In this book an attempt is made to discuss the application of physical chemistry to a number of general problems in inorganic chemistry, and it provides a very useful introduction to this approach. The topics covered range from atomic structure, valency, the solid state, and liquids and solutions, to the mechanism of inorganic reactions.

In particular, the chapters concerned with hydration and the nature of species in solution, and with the solid state are very useful, and the text contains some valuable tables of physical data. In general, the main points in the explanations of the chemistry of these systems are very effectively made, but occasionally anomalies are pointed out without any attempt to discuss the reason for this in terms of current theories. Thus in the section on electron affinity, the authors refer to the unexpected increase (in the electron affinity) from fluorine to chlorine, which also shows up in going from oxygen to sulphur, but they make no attempt to state the reason for this unexpected increase. This may cause some frustration for the average reader. However, this book will prove extremely useful as a summary text for many courses in inorganic chemistry. A very useful supplementary reading list is provided at the end of each chapter.

J. Lewis

SILICON CHEMISTRY

Volatile Silicon Compounds, *by E. A. V. Ebsworth. Pp. v+179. Pergamon Press, Oxford. 1963. 42s. net.*

The chemistry of Group IV is often presented as rather dull and obvious: silicon differs from carbon simply in having available *d* orbitals and a higher bond energy with oxygen, and the heavier elements show an increasingly stable +II state. The author gives us an account of silicon chemistry which shows that these properties are manifested in a much more subtle and complex way. Although *d*-orbital effects are important, they are often postulated on the basis of scanty evidence, and an outstanding feature of this monograph is the author's critical and discriminating approach to the physical and chemical evidence for these effects.

This book is a well-documented account of the simpler organic and inorganic molecules formed by silicon, covering principally those with silicon–hydrogen, silicon–halogen, silicon–nitrogen or silicon–oxygen bonds. Physical and chemical properties are discussed, and special attention is paid to infra-red and nuclear magnetic resonance studies. Shorter discussions of the corresponding germanium and tin compounds are also included.

The book does suffer, unfortunately, from an undue number of misprints and other printing blemishes. However, the author has produced a thorough, well-organized, and stimulating review which can be recommended to his fellow workers and to chemists in general

K. M. Mackay

CARBON CHEMISTRY

Chemistry of Carbon Compounds, *edited by E. H. Rodd, Vol. V:* **Miscellaneous, General Index.** *Pp. xviii+912. Elsevier Publishing Co., Amsterdam. 1962. £8 net.*

This work completes the series which was begun in 1951, and which now, having reached a total of nearly 8000 pages, stands alone as the most comprehensive general textbook of organic chemistry ever to appear in the English language. All those who have taken part in the enterprise must be able to look at the result of their labours with great satisfaction. The larger part (680 pages) of this final volume is devoted to a general index to the whole work. This has been compiled (by R. E. Fairbairn) with remarkable thoroughness; a spot check on a number of compounds mentioned in various parts of the complete text, but not emphasized in any way, showed that they all appeared, with the correct reference, in the index. The first 233 pages of the book contain four very interesting chapters on some topics that have undergone rapid expansion in recent years: isotopically labelled organic compounds (H. R. V. Arnstein); natural acetylenic compounds (L. Crombie); developments in carbohydrate chemistry (G. O. Aspinall); and the structure of proteins (H. D. Springall). These authors can be congratulated for ensuring that the high standard of 'Rodd' has been maintained to the end.

L. N. Owen

PAPER CHEMISTRY

Chimie appliquée à l'analyse des matériaux et produits cellulosiques (végétaux, pâtes cellulosiques, papiers), ***by Marcel Chêne. Pp. 304. J. B. Baillière et Fils, Paris. 1963. Fr. 25.50.***

The scope of this book is much wider than its title implies, and it would, in fact, be better described as an introduction to paper chemistry. It begins with a section on the chemistry of cellulose, hemicelluloses, lignin, and other constituents of plants, together with an account of the production and bleaching of cellulose pulps. The second section discusses the production of paper and board (without going into details about machinery), while the third gives a critical account of the analyses with which every paper chemist should be familiar. Detailed procedures are given only when there is no suitable French or American standard method to refer to. Finally there is an appendix containing basic ideas about the electronic theory of valency, ionic equilibria, redox potentials, and volumetric analysis, followed by tables of atomic weights, densities of various solutions, and certain dissociation constants.

Professor Chêne writes lucidly and, at times, with a dry humour; for example, in the middle of a discussion of the theory of rosin sizing we read '*certains auteurs ne se contentent pas de contredire les autres, il arrive qu'ils se contredisent eux-mêmes*'. However, in a number of places the discussion is so oversimplified that it is likely to mislead the uninitiated, for whom the book is primarily intended. Two examples may be cited: on page 50, pectic acid is said to be a complex galacturonic acid (having the same structure as cellulose with the CH_2OH groups replaced by CO_2H), and on pages 35 and 216 the essential difference between the action of cold concentrated alkalis and hot dilute alkalis on cellulose materials is not even hinted at. The number of misprints and mis-spellings of proper names is excessive.

T. P. Nevell

ZONE REFINING

Zone Melting of Organic Compounds, ***by E. F. G. Herington. Pp. viii+162. Blackwell Scientific Publications, Oxford. 1963. 35s. net.***

The author has studied zone melting for some years, and his wide experience is reflected in the pages of this useful book which collects the extant information. A short history of zone melting is followed by an account of the underlying theory, and this in turn by a discussion of important factors influencing the successful application of the technique and the available experimental methods. The author then considers problems of the storage and handling of pure organic compounds and the methods used to establish their purity, and gives an account of the several applications of zone melting in organic chemistry. He describes the usefulness of the techniques in the study of phase relationships of binary mixtures, and the text concludes with accounts of related separation techniques. Useful thermal data for some common organic compounds are appended to the text.

Zone refining may not be universally applicable to organic compounds, but the present text indicates its usefulness for the purification of many organic chemicals. The exceptional degree of purity attainable has particular significance in chemical standardization, where techniques providing compounds of undoubted purity are essential. This most useful monograph should interest not only research chemists, but also chemical technologists concerned with the production of ultra-pure chemicals. As an exposition of a basic process of chemical purification, it is a book to interest all chemists.

William I. Stephen

METEOROLOGY

Grosswetterkunde und Langfristige Witterungsvorhersage (Large-scale Meteorology and Long-range Forecasting of Weather Character), ***by Franz Baur. Pp. vii+91. Akademische Verlagsgesellschaft, Frankfurt am Main. 1963. DM 35.***

Professor Franz Baur is one of the few important world figures in the field of long-range weather forecasting. His interest has been continuous through half a century, and this record of five lectures delivered by him at a meeting in October 1961 is perhaps the best available introduction to his work.

In the first lecture it is concluded that periodicities and symmetry points are an unprofitable line, and that 'singularities', although statistically significant, are too irregular to form a basis for forecasting. In the second lecture Baur's well-known ideas on *Grosswetter* (large-scale weather) are outlined and a version of what elsewhere is called the 'index cycle' is put forward. Next, the physical reality of long-range weather variability, as something more than chance associations of weather types, is supported by statistical argument, and his method of picking out prediction rules of limited application is indicated. The fourth lecture deals mainly with sunspots, the influence of which is a notable feature of Baur's system. The fifth lecture illustrates the method by reference to the winter of 1961–62. There are a few other contributions, including a lecture by Attmannspacher 'On the connection between tropospheric and stratospheric circulation'.

As an outline of a legitimate empirical line of attack exploited by a wise and scholarly meteorologist, this slim, well-printed volume with its long English summaries will be valued widely, even if it does not seem to promise a basis for much further exploitation.

R. C. Sutcliffe

IMMUNOLOGY

Cell-bound Antibodies (Conference of the National Academy of Sciences—National Research Council—held May 1963). ***Proceedings edited by Bernard Amos and Hilary Koprowski. Pp. 134. The Wistar Institute Press, Philadelphia. 1963. $4.***

During the past few years, immunologists have begun to exploit more fully a variety of new techniques. In particular these have been applied to the enigmatic state of delayed hypersensitivity and other related conditions. As a result, new information is accumulating rapidly. The conference sponsors chose an opportune moment to discuss some of these new data.

One of the principal themes was the interactions between humoral antibodies and the cellular elements of an immune response, the emphasis being placed on the analysis of such interactions *in vitro*. The role of cytophilic antibody, a class of immunoglobulins having the property of becoming adsorbed on the surface of host cells, is discussed in an excellent article. Among some of the other subjects were leucocyte lysosomes, the opsonic effect of antibody in the phagocytosis of foreign cells, and the cytotoxic and protective functions of humoral antibody in allergic encephalomyelitis. The behaviour of sensitized lymphocytes in cultures of target tissue is described in two of the contributions, and there is an interesting article on the transfer of immunity to foreign skin grafts with extracts of sensitized lymph node cells.

The standard of all articles and the discussions is very high, and there can be little doubt that all who are interested in immunology will wish to study this book.

J. R. Batchelor

CELL DIFFERENTIATION

Biological Organization at the Cellular and Supercellular Level, ***edited by R. J. C. Harris. Pp. x+261. Academic Press, London and New York. 1963. 60s. net.***

Today there is a widespread feeling among biologists that the next main problem to be attacked is the way in which biological organization at a level higher than that of molecules is built up. This book reports a symposium sponsored by UNESCO at which this problem was discussed.

The book opens with a paper by Jacob and Monod. It has been known for some time that bacteria grown in the presence of a particular nutrient—for example lactose—may produce enzymes enabling them to utilize lactose, although they would not produce such enzymes in the absence of the sugar. Jacob and Monod have now elucidated the mechanism whereby the genes responsible for these enzymes are kept 'switched off' in the absence of lactose, and are 'switched on' in its presence. The work on bacteria may give us a clue as to how cell differentiation occurs in higher organisms. But unfortunately for our hopes of simplicity, other papers in the book, dealing with organisms more complex than bacteria, describe a number of phenomena which cannot easily be explained in terms of the mechanism discovered by Jacob and Monod.

The book is of too technical a nature to be easily followed by any but professional biologists. But for them it provides a stimulating series of reviews of recent work in a fascinating field of biology.

J. Maynard Smith

MORPHOGENESIS

Advances in Morphogenesis, *edited by M. Abercrombie and Jean Brachet, Vol. II. Pp. xi+388. Academic Press, New York and London. 1962. 89s. 6d. net.*

It is almost too easy to obtain experimental data on developing systems, but extremely difficult to plan and execute critical experiments. There is not yet a satisfactory theoretical framework or conceptual system in whose terms results can be discussed, and the result is the accumulation of a vast number of data on a wide variety of problems and systems whose significance is far from clear. The articles in this volume reflect these difficulties, but are invaluable for improving the position by providing reviews which bring together the accumulated data and ideas on specific topics. It is surprising that we have done without such a series for so long. The value of the articles would be improved if more attention were given to clearly stating the problems and summarizing the main results and conclusions. The articles cover: Morphogenesis in *Stentor;* Determination of symmetry in vertebrate eggs; Comparative biochemistry; Proteins in development; Histochemistry of regenerating limbs; Development of slime moulds; Early insect embryogenesis; Regeneration of planaria; and Experiments on aortic arch development.

L. Wolpert

COLOUR-CHANGE PHYSIOLOGY

Color Change Mechanisms of Cold-blooded Vertebrates, *by H. Waring. Pp. xii+266. Academic Press, New York and London. 1963. 76s. net.*

Workers in the field of colour-change physiology have long felt the need for a critical review of the nature of the mechanisms controlling the activity of the chromatophores of lower vertebrates. Professor Waring's book, much of it devoted to a description of, and comment on, the experiments made by Hogben and his co-workers on which Hogben based his hypothesis of a bihumoral control of lower vertebrate melanophores, does little to satisfy this need. At least one criticism of Hogben's bihumoral hypothesis has been missed, as have other relevant papers, notably that by Imai (1958). Much of the work of G. H. Parker's school has been only briefly mentioned, and this part of the book is as unbalanced in its way as was Parker's, published in 1948. The chapter on methods used in the assay of melanin-dispersing hormone (as Waring describes the hormone generally known as MSH, melanocyte-stimulating hormone) is useful, though over-detailed for a book addressed to senior undergraduate students. Chapters on the chemistry of melanin-dispersing hormones, the mechanism of the melanophore and the biological advantages of colour change are of interest. The book is well produced.

A. K. Kent

BIOCHEMISTRY

Proceedings of the Fifth International Congress of Biochemistry, Moscow, 1961. *General Editor N. M. Sissakian. Vol. I.* Biological Structure and Function at the Molecular Level. *V. A. Engelhardt (Ed.). Pp. xiii+261. Vol. II.* Functional Biochemistry of Cell Structures. *O. Lindberg (Ed.). Pp. xi+302. Vol. III.* Evolutionary Biochemistry. *A. I. Oparin (Ed.). Pp. xii+354. Vol. IV.* Molecular Basis of Enzyme Action and Inhibition. *P. A. E. Desnuelle (Ed.). Pp. xvi+343. Vol. V.* Intracellular Respiration: Phosphorylating and Non-Phosphorylating Oxidation Reactions. *E. C. Slater (Ed.). Pp. xiv+426. Vol. VI.* Mechanism of Photosynthesis. *H. Tamiya (Ed.). Pp. xi+385. Vol. VII.* Biosynthesis of Lipids. *G. Popják. (Ed.). Pp. xii+435. Vol. VIII.* Biochemical Principles of the Food Industry. *V. L. Kretovich and E. Pijanowski (Eds.). Pp. xvi+311. Vol. IX.* Transactions of the Plenary Sessions and Abstracts of Papers. *N. M. Sissakian (Ed.). Pp. viii+639. Pergamon Press, Oxford. 1963. Each volume £5 net. Complete set of nine volumes £42 net.*

These nine volumes comprise the Proceedings of the Fifth International Congress of Biochemistry, held in Moscow in August 1961, which was attended by over 5000 biochemists. The immense task of organizing this gigantic Congress fell to Russian biochemists, and of presenting the results in these volumes to the General Editor Academician N. M. Sissakian, to whom all biochemists owe a deep debt of gratitude. Thanks should also surely go to the planners and editors of the eight individual symposia. It would be invidious for a reviewer in a short review to pick out the stars in such a constellation. What is interesting is to note that enzyme actions are still given a large share of the Proceedings, in volumes IV, V, VI, and VII; but it is noticeable that the emphasis of biochemistry is widening, and includes the origins of life in Volume III, and biological structure and function in Volumes I and III. This is a trend which we can expect to continue, along with a deeper chemical analysis of the mechanism of enzyme action. Those who hope to find in these pages the very latest information on research in their own speciality will be disappointed. Some attempt has been made in several of the papers to add a few remarks embodying later work. However, the main value of the Congress publications is that they form a permanent record of biochemical research up to 1961 in a wide range of fields. In this sense they are a stocktaking of what is established, and a suggestion of where further work should proceed. Anyone wishing to study another field of biochemistry might well read the appropriate symposium to gain the necessary background in a short perusal. In such a rapidly developing field as present-day biochemistry, a survey of the whole is advisable every few years.

The final volume includes two distinguished lectures, by D. E. Green on 'The structure and function of subcellular particles', and by F. Šorm on 'Protein structure and function', together with accounts of the opening and closing sessions and the various individual communications.

The books are well produced and have adequate indexes.

Rudolph Peters

Comprehensive Biochemistry, *edited by Marcel Florkin and Elmer H. Stotz. Vol. IX.* Pyrrole Pigments, Isoprenoid Compounds, Phenolic Plant Constituents. *Pp. xvi+265. Elsevier Publishing Co., Amsterdam. 1963. 70s. net.*

This volume belongs to Section II of the work, the section that is concerned with the pure chemistry of biologically important compounds. It consists of a series of competent and well-written surveys of the chemistry of porphyrins, chlorophylls, bile pigments, and compounds derived from isoprene, and finally of a short review of our present knowledge of phenolic plant constituents. While the emphasis is on the structural chemistry, related physical chemical concepts are unavoidable, and many of the authors have also found it convenient to consider the biogenesis of the substances with which they are dealing. This leads sometimes to a somewhat scrappy treatment of aspects that one hopes will be fully covered in later volumes. The standard of the individual chapters is consistently high, and one is impressed by the amount of information compressed into 250 pages of print. The brevity, which is compatible with clarity, however, produces a book which is interesting and informative to the expert but might make rather heavy reading for the general biochemist. The book is well produced, the formulae are clear and the reviewer has found only an extremely small number of errors.

A. Neuberger

NUCLEOSIDES AND NUCLEOTIDES

The Chemistry of Nucleosides and Nucleotides, *by A. M. Michelson. Pp. ix+622. Academic Press, London and New York. 1963. £6 6s. net.*

This volume gives a detailed account of the chemistry of the nucleosides and the nucleotides, the latter including those large and most important groups, the nucleotide coenzymes and the nucleic acids. Biochemical and biological matters, too, are dealt with. No recent publication by a single author has attempted this, and the volume is therefore important. The attempt to deal with the wide ramifications of the subject has largely succeeded. The chapters on the structure, synthesis, and chemistry of nucleosides and mononucleotides are excellent. Reference is made throughout to very recent work.

The book has its faults. Of the large number of structural formulae, only a small fraction are numbered to give correspondence to the textual discussion. This must represent an enormous drawback for many readers. A considerable amount of unnecessary duplication is apparent, particularly in matters of chemical mechanism, where the author's views seem, at times, to be rather idiosyncratic.

The author, in his preoccupation with comprehensiveness has, it seems, lost the grand sweep of the subject. But this does not detract from the value the book must have for the many research workers interested in the biologically important organic phosphate esters.

There are very few errors detectable in the text. *D. M. Brown*

CEREBRAL CHEMISTRY

Chemical Exploration of the Brain (a Study of Cerebral Excitability and Ion Movement), *by Henry McIlwain. Pp. viii+207. Elsevier Publishing Co., Amsterdam. 1963. 40s. net.*

This book, which deals with the fundamental property of nervous tissue, excitability, should be of interest to anyone working in this field.

Professor McIlwain found that brain tissue suspended in salt solution was excitable to electrical stimulation, since it responded with increased oxygen and glucose utilization. No such increase occurred when the tissue was stimulated in the cold. Nevertheless the stimulation produced profound changes, for the tissue no longer responded to stimulation when returned to normal temperature.

This loss of excitability was imitated by certain constituents of brain extracts—basic proteins and histones—and restoration of excitability was brought about by others—neuraminic acid derivatives and gangliosides. The loss could therefore be explained by migration within the tissue, from the nuclei, of an inhibitory agent (a basic protein) and restoration of excitability by combination of a restoring agent (a ganglioside) with the inhibitory agent. One of the most profound responses of electrical stimulation is a fall of phosphocreatine, and the basic proteins were found to prevent the re-synthesis of this energy-rich phosphate. They do so by affecting the potassium movements. Normal tissue regains the potassium lost on stimulation within three minutes. This recovery is prevented by basic proteins and antagonized by gangliosides.

These and other new findings form the basis of McIlwain's theory of excitability. The last chapter deals with theoretical considerations of the active cation transport and the involvement of an adenosine triphosphate system in this transport.

W. S. Feldberg

PLANT VIROLOGY

Symptoms of Virus Diseases in Plants, *by L. Bos. Pp. 132. Centre for Agricultural Publications and Documentation, Wageningen, Netherlands. 1963. 23s. net.*

As the title implies, this book is concerned solely with symptomatology and, apart from a brief section devoted to virus particles and inclusion bodies as diagnostic aids, makes no reference to physico-chemical properties, classification, or vector relationships.

The author's aim has not been to produce a reference work for use in diagnosis, and indeed the absence of a general index would seriously limit its use for this purpose. His aim has been to describe and classify, by means of well-chosen examples, the types of histological and morphological deviations that result from virus infection and to rationalize the terminology used in their description. In this connection, the provision of an index of names of symptoms in five languages should prove invaluable to workers in a field that has now become international.

The book is clearly and concisely written in very readable English, and includes a bibliography of some 200 references. The 41 photographic illustrations are of quite exceptional quality. It is perhaps unfortunate that personal views are presented occasionally without adequate scientific documentation.—for example it is stated that the organoid malformations are practically all due to hormonal disturbances. Despite this, Dr Bos has made a valuable contribution to plant virus literature. *A. J. H. Carr*

MARINE BIOLOGY

Oceanography and Marine Biology, an Annual Review, *Vol. I, edited by Harold Barnes. Pp. 478. George Allen and Unwin Ltd, London. 1963. 75s. net.*

The contents of this volume are extensive and apparently disconnected: they include papers on pogonophora, mean sea level, microdistribution of plankton, underwater television, rhodophyta, molluscs, parasitic copepods, and the biogeography of the Australian coasts. Jerlov reviews the recent theoretical and instrumental advances in optical oceanography; the article is in the same class as his classical paper in the Reports of the Albatross Expedition. Arthur Lee describes the hydrography of the European arctic, and tabulates heat and water balances from various sources. He has also summarized the IGY work in the area, together with the ICES expedition on the Iceland Faroes Ridge. D. W. Hood's review of chemical oceanography tabulates some recent measurements of trace elements and summarizes the results of analysis and distribution of gases and nutrients. There is a controversial study of primary production ably written by Yentsch, and his summary of the large quantity of work on the radiocarbon technique is admirable. I also liked Stride's account of the geology of the continental shelves and Rossiter's brief and very clear review of tides. Dr Barnes is to be congratulated on his first volume. *D. H. Cushing*

The Natural History and Geography of the Antarctic Krill *(Euphausia Superba Dana)*, Discovery Reports, *Vol. XXXII*, issued by the National Institute of Oceanography, *by J. W. S. Marr. Pp. 33–464. University Press, Cambridge. 1962. £10 net.*

Krill lay their eggs in the surface waters, from which they sink to more than 1000 m. The depth of hatching is unknown, but the naupliar stages are found only below 500 m. Marr suggests that most eggs are laid in shelf waters, from which they are swept by cascading water into the ocean. His evidence does not support this idea, however, because the numbers of eggs found in oceanic waters are rather greater than those found in shelf waters.

The larval stages rise to the surface. The younger ones (calyptopids) appear to migrate diurnally between 250 m and the surface, but the older ones (furcilia) do not do so. Because the depths of capture of the younger stages between 1800 hours and 2200 hours are very variable, Marr believes that the migration is an artefact caused by their ascent from the depth of hatching. But at the season of capture there would be considerable differences in the times of nightfall, and hence variations in the depths of migrants. However, if the apparent vertical migration is really a developmental ascent, then the animals must arrive at the surface at night and metamorphose immediately. The conflict between vertical migration and developmental ascent is not resolved.

There is a good account of the westward drift of krill in the Weddell Sea. There is considerable discussion of patchiness. Large krill dodge the larger nets, and perhaps both large and small krill dodge the small ones. The work contains many references.

D. H. Cushing

PHARMACOLOGY

British Pharmacopoeia 1963. *Pp. xxviii+1210. Official from January 1st 1964. Published for the General Medical Council by the Pharmaceutical Press, London. 1963. £5 net; and* British Pharmaceutical Codex 1963, *published by direction of the Council of the Pharmaceutical Society of Great Britain. Pp. xxxvi+1433. Pharmaceutical Press, London, 1963. £5 5s. net.*

The first stage in the official recognition of a new drug which has passed appropriate laboratory and clinical tests is an approved name from the British Pharmacopoeia Commission of the General Medical Council. If it proves itself, it will enter the Codex to the British Pharmacopoeia, a surprisingly readable reference work giving the properties, actions, and uses of each drug, and much supplementary pharmaceutical information. For remedies of substantial, general, and well-proved worth, there is the ultimate canonization of entry in the British Pharmacopoeia itself. This sets a standard for preparation of the drugs in it, a standard with statutory authority for those drugs to which the Therapeutic Substances Act in Britain applies. The Pharmacopoeia is made up largely of technical detail, but Appendix xv is a brief but excellent monograph on statistical method, required for biological assay of about 30 out of around 900 substances, mostly antitoxins and hormones which still baffle the chemist.

The new B.P., apart from many minor revisions, features 260 new remedies and 105 deletions, a rate of change of around 6 per cent per annum. The additions reflect important advances: new tranquillizers, antibiotics, and anti-cancer drugs, and a new radio-isotope. Tinctures and infusions, once a dominant feature, now number only 16. In the Codex, boric acid goes, but spectrophotometry and tests for aerosol propellants enter. Investigation into antiseptic dressings and eyedrops is called for. Should a surgical dressing contain an antiseptic? It seems a small question: yet cuts and burns, suffered by millions, are too easily neglected in favour of some dramatic rarity.

A topical feature is the conversion to the metric system, complete in the B.P., not quite complete in the Codex, reflecting the increasingly scientific aspect of medicine, and the increasing use of pure substances.

The two volumes represent an immense amount of work, with close co-operation between the British Pharmacopoeia Commission, the Pharmaceutical Society (of Britain), who are responsible for the issue of the Codex, and the medical profession. They command gratitude, and also admiration for the quality of the printing, maintained high in the face of rapid change.

W. D. M. Paton

CONSERVATION

Recent Advances in Conservation, ***edited by G. Thomson. Pp. xv+224. Butterworths, London. 1963. £5 5s. net.***

This handsome volume has pretensions to be taken as seriously as its illustrious models in physics and biology. The International Institute for Conservation was barely ten years old when the Rome Conference, reported in this volume, took place and the contributions selected for publication are not all strictly advances nor all about conservation. However, the physics and biology which are used have the full traditions of those disciplines behind them, and some recent developments have been applied with astounding speed; already there is an obvious need for review papers. But perhaps most significant is the general insistence on scrupulous examination before any conservation is attempted; and nearly a tenth of the space is devoted to training.

Although the contents are still clearly divided between 'Pictures' and 'The rest' it is good to see some firm attempts at integration, as in the sections on museum climate, methods of analysis, and consolidation of friable materials. Such hardy perennials as metallic objects and wood panel paintings apart, specific attention is devoted to textiles, documents (the longest section), and to fresco transfer, these latter two sections by way of a tribute to intensive study in the host country. Detailed descriptions, with practical details, range from massive monumental bronzes to *millefiori* fragments. In the latter, shown, with some other items, in excellent colour, individual particles of opacifying agents were examined by electron-beam X-ray microanalysis. This method, featured in different ways in three sections, has shown immense promise here, as elsewhere. Its use in studying directly the 'chemical shape' of pigment particles in paintings offers attractive scope for advance. Meanwhile, conventional X-ray fluorescence analysis is being adapted to non-destructive work on large paintings. To talk in this way is almost to shrug off the enormous advances of the first ten years of the International Institute for Conservation. All credit to organizers and editor, and far more power to their elbows.

L. Biek

BIOGRAPHY

Biographical Memoirs of Fellows of the Royal Society, ***Vol. IX. Pp. 321. The Royal Society, London. 1963. £2 net.***

At the best of times, obituary notices make somewhat melancholy reading. Nevertheless, where both the subject and his work are familiar to the reader they have a very considerable interest. The volumes of detailed biographies of deceased Fellows of the Royal Society, of which this is the ninth, are—and will increasingly become—important works of reference. All are prepared by distinguished writers who not only have themselves a profound knowledge of the late Fellow's own field of research, and so can critically assess it, but have been prepared to go to immense trouble to seek out all the available biographical information. There thus emerges a very full picture indeed o each man and his work. There is no fixed length, but the average is around 8000 words; there is a photograph of each subject. Those whose lives are recorded in the present volume include Niels Bohr (by Sir John Cockcroft), R. A. Fisher (F. Yates and K. Mather), John Read (E. L. Hirst), M. W. Travers (C. E. H. Bawn), and G. M. Trevelyan (Lord Adrian).

Trevor I. Williams

HISTORY OF SCIENCE

The Collected Papers of Lord Rutherford of Nelson, under the scientific direction of Sir James Chadwick, ***Volume II.*** **Manchester.** ***Pp. 590. George Allen and Unwin Ltd, London. 1963. £5 5s. net.***

This, the second volume of Lord Rutherford's collected papers, is devoted to the period of his professorship at Manchester. It was in the summer of 1907 that he left McGill for the English university, and the Nobel Prize was awarded to him in the following year for the work which he had done in Canada in establishing the fundamental laws of radioactive transmutation. It was therefore as a world-famous figure in physics that he took up the Manchester appointment.

The keynote of his work at Manchester was the properties of the alpha particle and its use as a probe for investigating problems of atomic structure. Possibly the most famous paper in the volume is that published in 1911 on 'The Scattering of the α and β Rays and the Structure of the Atom', in which, as a result of consideration of the large-angle scattering of alpha particles, the nuclear structure was first put forward. The jubilee of this paper was the occasion of a great Rutherford Jubilee International Conference held at Manchester in 1961 and, incidentally, of a leading article in ENDEAVOUR on the jubilee of the nuclear atom.

One of the first papers in the volume, written in collaboration with Geiger, deals with the fundamental phenomena employed in the famous Geiger counter and another deals with the scintillation method, used in the experiments on which the nuclear structure was based. The last paper in the volume is concerned with the first demonstration of the disruption of a nucleus, by the impact of an alpha particle. Although the alpha particle dominates in this volume, there are many papers of great importance on beta particles and gamma rays.

The volume contains a preface by Sir James Chadwick, a most lively and illuminating article by Professor N. Feather on 'Rutherford at Manchester' and reminiscences of Rutherford at Manchester by H. Geiger, E. N. da C. Andrade, and A. B. Wood. Well-chosen plates add to the interest of the collection.

E. N. da C. Andrade

HISTORY OF BIOLOGY

The Science of Life: a Picture History of Biology, ***by G. Rattray Taylor. Pp. 368. Thames and Hudson, London. 1963. 42s. net.***

This book contains a great deal of information intermingled with some startling misinformation. It has, however, no pretensions to exact scholarship, and must be considered as a work of popularization. As such, it seems to me not entirely satisfactory.

The author seems to have set out in the belief that the history of the biological sciences is easy enough to understand and needs only to be made interesting; in fact it is interesting in itself, but requires a great deal of elucidation. In both the text and the illustrations a lot of space—which could well have been devoted to explanatory text and to more relevant illustrations—seems to have been wasted on purely extraneous matter. There are too many anecdotes, too many pictures of persons and places, and too many value-judgments, not all of which escape crudity.

The general effect is moreover breathless and jolting. It is like being taken round the National Gallery by a not always reliable guide on a motor scooter, with frequent but only momentary stoppings and startings. However, if the alternative is not to see the pictures at all, perhaps there is something to be said for the method. The visitor gets some idea of what is in the collection and may perhaps return to get a more leisurely view.

J. S. Wilkie

90

ENDEAVOUR

VOLUME XXIII SEPTEMBER 1964

A review of the progress of science, published in five languages

Editor TREVOR I. WILLIAMS, M.A., B.Sc., D.PHIL., F.R.I.C.
Foreign Editor G. N. J. BECK, B.Sc., A.M.I.E.E.
Assistant Editor W. A. OSMAN, B.Sc., M.Sc., A.R.C.S.

Contents

371.32-77
114 Editorial:

537.312.62:538.66
115 Superconducting magnets
J. E. Kunzler, B.S., Ph.D.

541.183.5:553.066
122 Molecular sieves
R. M. Barrer, D.Sc., Sc.D., F.R.S.

539.166
131 Some applications of the Mössbauer effect
P. B. Moon, M.A., Ph.D., F.R.S., and D. A. O'Connor, B.Sc.

551.578.11:532.69
136 The collision, coalescence, and disruption of drops
B. J. Mason, D.Sc.

543.544.25:53.082.8
142 Physical measurement by gas chromatography
J. H. Purnell, M.A., B.Sc., Ph.D.

591.51:159.929
148 The learning of discriminations by animals
N. S. Sutherland, M.A., D.Phil.

581.174:547.458
153 ✓ Structural plant polysaccharides
R. D. Preston, Ph.D., D.Sc., F.Inst.P., F.L.S., F.I.W.Sc., F.R.S.

621.391:681.142
160 The language of machines
Sir Edward Bullard, M.A., Sc.D., F.R.S.

165 Book reviews

Some books received (cover iii)

Communications to the Editor should be addressed to South Block, Thames House, Millbank, London S.W.1 Telephone: Tate Gallery 2181

Books for students in developing countries

There is no easy road to a higher standard of living for the world's underdeveloped countries. The complexity of the social, economic, political, and other factors involved is such that progress is necessarily slow and the best route a subject for controversy. Nevertheless, there is general agreement about certain conditions that are a prerequisite for progress, and high among these is an adequate system of education at all levels, especially in science and technology. Without this, the trained personnel necessary for establishing and running a modern state cannot be made available. This is a lesson that western states have in the past had to learn by hard experience. A century ago, the Great Exhibition of 1851 seemed to set the seal on Britain's industrial supremacy, but discerning observers realized that this was not the beginning but the end of an era. At the Paris Exhibition of 1867, Britain won barely a dozen awards, and the decline was directly attributed to the fact that whereas other countries had actively developed their state educational systems in accordance with the needs of the day, Britain had not. Gradually, the lost ground has been recovered and the recent publication of the Robbins Report—presaging an unparalleled expansion of Britain's system of higher education—is a reminder of the great importance attached to higher education in the modern state.

To establish an advanced system of higher education many conditions must be fulfilled, and among the most important of these is the ready availability of first-rate books. To the affluent nations of the world this presents no real problem; generally speaking, the studies of western students are not limited by lack of books. In the developing countries the situation is totally different. National poverty is acutely reflected at student level, and in many countries of Asia and Africa it is a fact that the cost of textbooks not only prevents the student building up his own working collection, but even limits their availability to him through libraries.

Such considerations led to the establishment by Britain in 1960 of the Low-priced Books Scheme. Briefly, this is a scheme by which British publishers, aided by a government subsidy made available through the Central Office of Information, provides books to a number of Asian and African countries at a price equivalent to roughly one-third, or sometimes less, of that ruling in the home market. While the subsidy is the main reason for the very low price, some reduction is also effected by slight economies in paper, the use where possible of paper covers in place of hard bindings, and above all by very large print orders. The last may be ten or more times greater—sometimes a hundred times greater—than would have been possible at the original price. In all cases, however, the books are full and unabridged in respect of both text and illustrations. The books appear under the imprint of the English Language Book Society—a collective imprint adopted by the participating publishers, now numbering about fifty—though they are distributed by the individual publishers.

Clearly, a first requirement for the success of such a scheme is that there should be a most careful choice of the books included in it. From the outset, this has been entrusted to a panel of experts independent of Government and under the chairmanship of Mr A. L. P. Norrington, President of Trinity College, Oxford, and a former Vice-Chancellor of the University. Apart from his high academic standing, Mr. Norrington has had great experience of book publishing. He was formerly Secretary to the Delegates of the Oxford University Press, and has first-hand knowledge of the book needs of many of the territories included in the scheme. No book is included in the scheme unless the committee is satisfied that it is authoritative, as far as possible up-to-date, and suitable for use by overseas students. This last criterion is, of course, of the first importance and, to see that it is satisfied, advice is sought from many sources, including experts in the countries concerned. In this respect, the advice of the British Council has been invaluable. In India, the largest single purchaser of the books, the choice of textbooks is made in consultation with her Ministry of Education and University Grants Committee.

The books included in the scheme are all of an essentially educational character. There is heavy emphasis on science, engineering, and mathematics as representing fields whose development is of particular importance to the countries concerned. But a number of other subjects are represented among the college textbooks, now numbering more than eighty titles; they include economics, business administration, law, and history. The college textbooks, however, represent only about half of the two and a quarter million individual volumes so far produced. The *Aids to English* series is an obvious complement to all the others, for by no means all the students are fluent in the language which is their main medium of instruction. *The Teacher's Bookshelf* offers a series of first-rate texts on teaching methods, with special reference to the teaching of English. The *Science Today* series is designed to meet the needs of the non-specialist reader for simple accounts of the progress of modern science, and *The Modern World* concerns itself at the same level with industry and technology. *The Practical Books* deal with such topics as the use of the slide-rule, book-keeping, and typewriting. Very recently, it has been decided to extend the scheme to medical books, and to books catering particularly for the needs of students in technical colleges.

The developing countries must concentrate on matters of direct practical application, but altogether to exclude more general reading is clearly undesirable. For this reason, the *English Library* series was introduced at an early stage, making available a number of general books in good English prose. To mark the 400th anniversary of Shakespeare's birth a complete Shakespeare—the cheapest in the world—has just been published, and the demand has been remarkably large.

It is recognized that while many countries are today almost completely dependent on imported textbooks, their own publishers are understandably anxious to supply their national needs. The English Language Book Society has no wish to discourage this, and books directly competitive with locally published works are avoided where possible. British publishers have actively collaborated in fostering this ambition by accepting as trainees representatives of overseas publishing houses anxious to fulfil the educational needs of their own countries.

The magnitude and urgency of the need that the scheme seeks to meet is not easily appreciated in more fortunately placed countries. Dr V. S. Jha, a former Director of the Commonwealth Education Liaison Unit and sometime Vice-Chancellor of Banaras University, has stated it cogently. 'I have known the hunger for books; the savage rage which an ambitious young intellectual feels when penury keeps him away from books. . . . People will starve to secure education for their children. . . . Britain's scheme for the supply of textbooks goes straight to him or her who most needs it—the young student.'

If the subsidy is to be used to the best advantage, the range of such a scheme must naturally be limited to the countries where the need is greatest. Its scope has, in fact, been considerably enlarged since its inception and it now extends to thirty-nine states, in Asia and Africa. Within these countries, books published under the imprint of the English Language Book Society can be purchased through ordinary trade channels.

Superconducting magnets

J. E. Kunzler

The discovery of superconductors by Kamerlingh Onnes in 1911 was immediately followed by the suggestion that they might be used for generating intense magnetic fields with a relatively low expenditure of energy. Practical difficulties, particularly the fact that the earliest materials lost their superconducting properties in quite small magnetic fields, resulted in no useful progress then being made. Quite recently, however, the discovery of superconductors having different properties has made feasible superconducting magnets producing a field of as much as 100 kilogauss.

In theory, electromagnets can be made to produce fields of any desired magnitude simply by increasing the current through the coils of a solenoid. In fact, this conventional method presents very great practical difficulties, even for fields of less than 100 kG (kilogauss), mainly because of the problems of supplying the large amount of power and of dissipating the great quantity of heat—proportional to the square of the current—generated in the field coils. For this reason there has long been interest in the possibility of using superconductors for constructing the coils of electromagnets, since in the ideal case the heat loss due to the i^2R factor will be nil. Early attempts were unsuccessful because the materials used lost their superconductivity in quite low magnetic fields, but by 1961 alloys were discovered that did not possess this defect. Since then progress has been rapid.

The principles governing the construction of superconducting magnets, and the practical significance of their availability, are most easily appreciated by first briefly considering the limitations of large permanent magnets and large electromagnets.

Conventional high-field magnets

Under ordinary conditions even the best conductors, such as copper, have such high resistivities that the generation of steady magnetic fields as large as 100 kG is a formidable and expensive undertaking.

Magnetic fields of a few kilogauss, operative in volumes of a few tens of cubic centimetres, have been available for decades by using permanent magnets—a good horseshoe magnet has a field strength of about 1 kG. This field results from the aligning of part of the magnetic domains in an iron alloy. Fields as large as 20–30 kG, operative in rather small volumes, can be achieved in electromagnets by winding copper wire on an iron core. The relatively small magnetic field produced by passing a direct current through the windings causes the magnetic domains in the iron to align themselves parallel to this field and thus produce a much larger total field. The upper limit of the fields produced with iron-cored electromagnets is determined by the saturation magnetization of the core when all the domains are aligned; for the most favourable materials it is about 30 kG.

To produce magnetic fields higher than about 30 kG, one resorts to a 'brute force' technique. In the case of an infinitely long air-cored solenoid, the field strength in gauss is

$$H = 4\pi ni/10 \qquad \ldots(1)$$

where n is the number of turns of the windings per cm of length of solenoid, and i is the current in amperes through the windings. If the solenoid is at least as long as its diameter, equation (1) can be expressed, usually with an error of not more than 10 per cent, as

$$H \simeq ni \qquad \ldots(2)$$

It can be shown that equation (2) is equivalent to

$$H \simeq kjt \qquad \ldots(3)$$

where k is the fraction of the volume of the layers of windings that is occupied by the conductor (in the ideal case, $k=1$), j is the current density in amp/cm², and t is the thickness in cm of the windings.

One might expect that any desired field could be attained simply by increasing the current. To a certain degree this is true, but the power required increases rapidly—faster than the square of the field, because heat removal and mechanical constraint problems generally require that the physical size of the magnet and hence its effective electrical resistance also be increased. These problems become serious even before a field of 100 kG is attained. For example, a radial force exists at the mid-section of the solenoid; it arises from a 'magnetic pressure' that varies as the square of the field strength and is about 6000 lb/in² at 100 kG. There are perhaps one or two dozen magnets in existence that use normal conductors for the coils and are capable of producing fields of 100 kG. Each requires about 2000 kilowatts of power, cooling equivalent to about 1000 gallons of water per minute, and a capital outlay of several hundred thousand dollars. The National Magnet Laboratory at Cambridge, Massachusetts, plans to include a 250 kG magnet of exceptionally efficient design in their facilities. This multi-million dollar installation will require about 8000 kW of power, and its cooling will raise the temperature of the Charles River nearly 0·5°C. The most powerful magnet of this kind known to the author is the 150 kG magnet at the Naval Research Laboratories in Washington, D.C.

The enormous power requirements for high-field solenoids, such as those described above, are a result of the resistivity of the conductors used in the windings. However, once a magnetic field is established, in principle no power is required to sustain it (for example, no power is required to sustain the field of a permanent magnet). The electrical energy is converted to heat because of i^2R losses. Since one of the properties of superconductors is lack of resistance, the potential advantage of using them in magnets is obvious. Other specific properties are, however, necessary for this application. It is also necessary for the superconductor to have zero, or negligibly small, resistance in the highest magnetic field to be generated by the magnet and while carrying a current of the requisite density. Current-density requirements are dictated by equation (3), from which it is apparent that a current density of at least 10^4 amp/cm² would

John E. Kunzler, *B.S., Ph.D.*
Was born in 1923 and educated at the Universities of Utah and California (Berkeley). He joined Bell Telephone Laboratories in 1952 and is now Head of the Metal Physics Research Department. He has specialized in low-temperature solid-state research, including the investigation of the electrical, thermal, and magnetic properties of solids. He is one of the research scientists whose work led to the practical realization of a superconducting magnet.

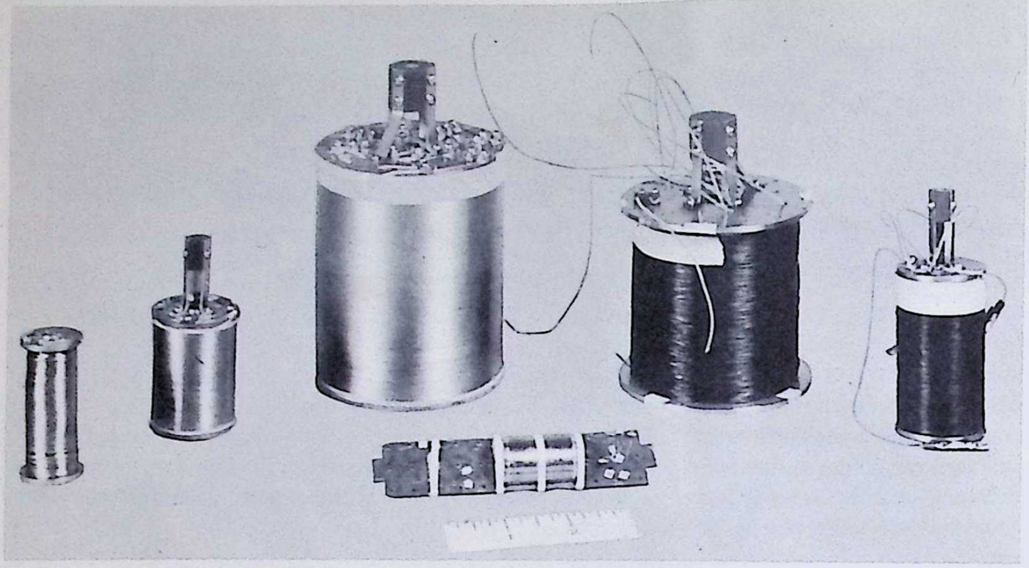


Figure 1 Photograph of early superconducting magnets constructed from Mo-Re (foreground), Nb-Zr (right), and Nb_3Sn-cored 'wire' (left). The scale is 3 in. long.

be required to generate a field of 100 kG if an ideal packing factor ($k=1$) were achieved, and if the thickness of windings were restricted to 10 cm. For the extremely unfavourable case of a packing factor of 0·1, a current density of 10^5 amp/cm^2 would be required. The superconducting materials now available for the construction of magnets satisfy these general requirements.

Superconductors for magnet construction

Recent developments in this field stem from the unexpected discovery in the Bell Telephone Laboratories [1] that even in high magnetic fields, certain superconductors, such as Nb_3Sn, could carry current densities as large as 10^5 amp/cm^2. Not for the first time, hopes were raised that powerful magnetic fields could be generated for a relatively small expenditure of energy. Immediately after the discovery of superconductivity [2] at Leiden University in 1911, Kamerlingh Onnes recognized the possible advantages of superconductors for this purpose. Although some of his initial observations, such as the measurement of critical current, were most encouraging [3], his hopes were dashed when he found that a magnetic field of only a few hundred gauss destroyed the superconductivity of the materials that he had been investigating (Hg, Pb, Sn) [4]. In 1930, W. J. de Haas and J. Voogd, also of Leiden, found that Pb-Bi alloys remained superconducting in moderate fields and suggested that superconducting magnets capable of producing fields as large as about 20 kG should be possible [5]. Although the idea was abandoned after further observations proved discouraging [6], it is now known that a superconducting magnet having capabilities approaching their expectations is in fact possible [7]. It was perhaps unfortunate for them that a technological need, such as that which played an important role in the ultimate achievement of practical superconducting magnets, did not then exist. Had a superconducting magnet been successfully constructed at that time, it would probably have speeded up the technology of superconducting magnets by at least a decade or two. It is hard to envisage the effect such an achievement might have had on the technological development of cryogenics and on our fundamental understanding of the phenomena of superconductivity.

There were three major factors in the recent series of events that resulted in the practical realization of high-field superconducting magnets. Firstly, cryogenic technology has advanced so much in the past few decades that low-temperature work is now feasible in many laboratories instead of being restricted to a small number of specially equipped ones. Secondly, a wealth of information concerning superconductivity and superconducting materials has resulted from the extensive studies of B. T. Matthias of Bell Telephone Laboratories and J. K. Hulm of Westinghouse Research Laboratories, and their associates. Thirdly, but by no means least, there is now a real need for superconducting magnets. For example, S. H. Autler at the Lincoln Laboratories needed magnets capable of field strengths of a few thousand gauss for the operation of solid-state microwave masers. Since the maser itself required refrigeration at liquid helium temperature, the possibility of using a superconducting magnet was particularly attractive. Autler's success in constructing a niobium-wound solenoid that would produce a field of 4·3 kG [8] encouraged R. Kompfner of Bell Telephone Laboratories to point out to several of his associates the potential value of superconducting magnets capable of producing somewhat higher fields for use in microwave electronics, in which he was interested. In both cases the need was sufficient to keep research going, even though only a few years earlier it had been generally believed that superconducting magnets capable of as much as a few kilogauss were not possible [9]. It must not be forgotten that a significant step forward had been made by G. B. Yntema, who in a short note [10] in 1955 reported the operation of an iron-core electromagnet having superconducting windings. This achievement did not attract wide attention, however, probably because there was uncertainty as to the field strength at the windings in that configuration.

By present standards, the magnetic fields sought for superconducting maser magnets were quite modest; 10 kG was hoped for, and fields as large as 15 kG were talked about. One of the first materials to be investigated at Bell Telephone Laboratories for the windings was an alloy of molybdenum and rhenium. Matthias suggested this material, not because it was expected to provide the highest possible critical field, but because its critical temperature, T_c (below which the resistance is zero), is among the highest ($\sim 10°K$) of the ductile materials, and was thus expected to be one of the best of the potential materials from the point of view of fabrication into wire. Using Mo-Re, it proved possible to construct a superconducting magnet (figure 1) that produced a field exceeding 15 kG [11], thus coming up to the most optimistic expectations.

By this time (1960), it seemed likely that there would be materials permitting fields of at least 20 kG, and perhaps as much as 30. However, the obvious materials having high critical temperatures (for example Nb_3Sn, $T_c = 18°K$; V_3Si, $T_c = 17°K$; and V_3Ga, $T_c = 16°K$) were brittle compared to Mo-Re, and most of the attention was turned to other ductile materials, such as Nb-Zr and Nb-Ti. Even after Nb_3Sn had been tested in fields as large as 18 kG and had been found to be superconducting while carrying large currents, no attempt was made to test the current-density properties of the material in the Bitter magnet—then capable of fields up to 88 kG—until a procedure of fabricating Nb_3Sn 'wire' had been developed [1], This was because the expected modest gain in field strength seemed hardly worth pursuing unless a method of making magnets from the material were available.

The unexpected observation that small samples of the brittle Nb_3Sn sustained a current density of 3000 amp/cm² in a field of 88 kG was followed within a few days by the equally surprising observation that the cored 'wire' (figure 6), when properly prepared, sustained a current density in excess of 100 000 amp/cm² in a magnetic field of 88 kG [1]. After the observations had been confirmed, it was apparent that high-field superconducting magnets would at last become a reality. Results showing that high-field superconductivity was not unique to Nb_3Sn were soon to follow. Nb-Zr [7, 12, 13], Nb-Ti [7, 12], V_3Ga [14], and V_3Si [14] were among the first materials found to have potential applications as magnet materials.

Since 1961, practical superconducting magnets capable of giving fields as high as about 70 kG have been constructed of Nb-Zr alloys and of Nb_3Sn. A feasibility model constructed of Nb_3Sn by workers at the General Electric Research Laboratories has operated at a field of 100 kG [15]. It is expected that operational magnets capable of fields in excess of 100 kG will be available soon.

Superconductivity

The history of our understanding of the physics of high-field superconductors is even more involved than that of the developments leading to superconducting magnets. Only a brief summary of our present understanding, and of some of the essential events leading to it, can be included here. Much of our new knowledge has evolved since the initial observations of the high-field properties of Nb_3Sn, which were not only unexpected but not easily explainable in terms of theory. The basic theory did in fact exist, but it was not widely understood. Our ideas have since changed considerably, and more advances are to be expected in the future.

Superconductors can conveniently be divided into two classes: Type I (also known as nearly ideal, or superconductors of the first kind) and Type II (superconductors of the second kind). High-field superconductors are of Type II; however, they must be in a physically inhomogeneous structural state before they can sustain the high current densities required for magnets. Type I superconductors have also been known as 'soft' and Type II, when in a state capable of supporting a transport current, as 'hard'; but these terms are ambiguous. Both Type I and Type II are characterized by a transition or critical temperature below which the resistance is zero, or at least too small to have been measured so far. In the upper part of figure 2, the variation of resistance with temperature of a superconductor is compared with the behaviour of a typical non-superconducting metal and an ideal metal (an infinitely large, pure, perfect single crystal) which has zero resistivity at 0°K. About two dozen elements and well over two hundred alloys and compounds have been found to be superconductors. Their critical temperatures range from a few thousandths of a degree Kelvin to 18°K for Nb_3Sn, discovered by Matthias and his co-workers [16] in 1954.

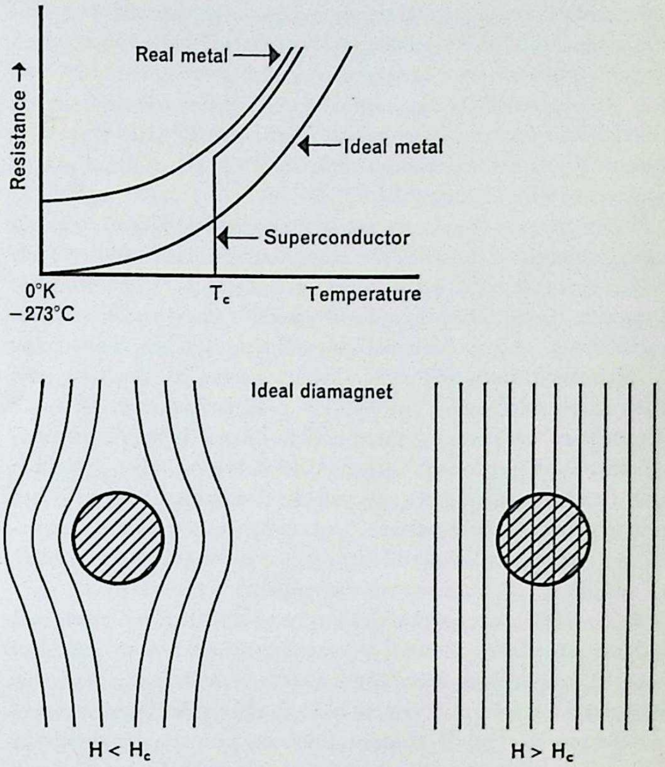


Figure 2 Loss of resistance of a superconductor at its critical temperature contrasted with resistance-temperature characteristics of a real and an ideal metal (top). The bottom of the figure shows the reversible diamagnetic behaviour of a nearly ideal Type I superconductor in a magnetic field.

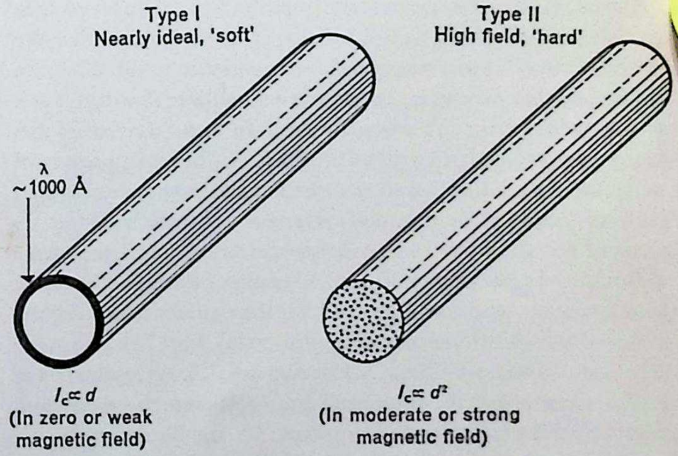


Figure 3 Critical-current characteristics of superconducting wire. The current is carried in a surface layer of Type I superconductors, and throughout the bulk of the material in Type II.

Type I superconductors are nearly ideal diamagnets. That is, a weak magnetic field is excluded from them—except for a thin layer on the surface about 1000 Å thick, known as the penetration depth (figure 2, bottom, left). At a field strength equal to the critical field strength, usually less than a few hundred gauss, the field suddenly penetrates the entire sample and at higher fields the material behaves in the same way as typical non-superconducting metals do. This effect is nearly reversible, and when the magnetic field is decreased to the value of the critical field, the magnetic field is expelled from the sample. This is known as the Meissner-Ochsenfeld effect [17], after its discoverers. Current flow in Type I superconductors is restricted to the thin 'penetration depth' layer, and thus the critical current—that is, the maximum possible while the material still remains superconducting—of a wire varies according to its diameter, as is illustrated in figure 3 (left). Most Type I superconductors are pure elements, such as lead, mercury, and tin, which are often mechanically soft; hence the designation 'soft'. Their properties are nearly ideal, and they can be readily treated by thermodynamics and the BCS (Bardeen, Cooper, and Scrieffer) Theory [18].

Type II superconductors are more complicated and less well understood. The 'ideal' Type II superconductor cannot sustain a transport current in a high field, such as is required for magnets, even though in some cases it may remain superconducting in very high fields—well over 200 kG. They show a Meissner-Ochsenfeld effect in the sense that the magnetic field is expelled when an applied field is removed. In weak fields, they shield in the same way as does a Type I superconductor. However, once flux penetration begins, it is not sudden but continues to develop as the field increases. Some other properties of nearly 'ideal' Type II superconductors are discussed below, and those of high-field superconductors capable of sustaining high current densities in the next section.

It has been known for many decades that some degree of superconductivity persists in certain alloys up to fields of several tens of kilogauss. It was also known from the extensive theoretical work of F. London [19] that if a Type I superconductor was made thinner than its penetration depth, it could be made to remain superconducting in fields many times its thermodynamic critical field. This happens because the superconductor no longer has to do the entire work ($H^2V/8\pi$, where H is the field and V the volume) of excluding the applied magnetic field. Consequently, the free energy of a thin sample in the magnetic field is reduced relative to that of the bulk material. Until a few years ago, it was generally believed that the persistence of superconductivity at fields well above the critical field of Type I superconductors was due to the existence in them of thin filaments, perhaps in a mesh-like structure [20], which would be stable once formed. In fact, the initial explanation for the high critical field of Nb_3Sn was in terms of such filaments [1], which were thought to be perhaps associated with dislocations [7, 21, 22]. However, it was difficult to account for the very large current densities, and even more difficult to reconcile the idea of filaments with the high-field heat capacity data for V_3Ga, which showed that most of the material was still superconducting at 70 kG [23]. At about the time that this became clear, it also appeared, largely due to the initial efforts of B. B. Goodman [24, 25], that the combined theories of several Russian workers, V. L. Ginsburg, L. D. Landau, A. A. Abrikosov, and L. P. Gor'kov [26], accounted for Type II superconductivity; this became known as the GLAG theory. They showed that if the effective coherence length, ξ_{eff} (essentially the average distance over which an electron remains superconducting or paired in the BCS theory), is shorter than the penetration depth, λ, the superconductor could spontaneously break up into alternate superconducting and normal regions (with a thickness less than the penetration depth) and that the surface free energy between these would have a negative value. In such circumstances, a large portion of the material is superconducting; it is stable and can be treated thermodynamically. It was also shown that the effective coherence length is controlled by the mean free path of the electron, l, when l is appreciably less than ξ_0, the coherence length of the pure material in bulk.

If ξ_{eff} is greater than $\sqrt{2}\lambda$, a superconductor is of Type I; if ξ_{eff} is less than $\sqrt{2}\lambda$, it is of Type II. Although it is necessary also to consider the electronic properties of a superconductor in making any generalization, it turns out that most pure metallic superconductors are of Type I, and most alloys—and intermetallic compounds, but for less obvious reasons—are of type II. This happens because the electron mean free path is relatively long for pure metals at low temperature and is often several orders of magnitude shorter for alloys, owing to the scattering of electrons by the impurity atoms.

High-field superconductors capable of high current densities

An ideal Type II superconductor in a high magnetic field cannot support a transport current—a current from an external source. The Lorentz force ($\mathbf{i}\times\mathbf{H}$) attempts to sweep the lines of magnetic flux from the material and it is 'driven normal', that is, it is no longer superconducting. However, most type II superconductors are not ideal and do support some transport current; in fact, it is difficult to prepare a nearly ideal type II superconductor. The reason for this is that the lines of flux are 'pinned' by defects—chemical, structural, or both—in the material. This phenomenon opposes the Lorentz force and hence permits a transport current. The maximum permissible current density, or critical current density, is determined by the strength of the 'pinning' and the number of 'pinning' sites. As the current density in a type II superconductor approaches its critical value, a 'creep' in flux is observed. This has been studied extensively by Y. B. Kim and his associates [27], and has been treated theoretically in terms of an activation process by P. W. Anderson [28].

Apart from making a transport current possible, there are two other consequences of flux 'pinning' and flux 'creep'. Firstly, the transition between the superconducting and non-superconducting state is not well defined. Secondly, the magnetization of these materials is not reversible, that is, they exhibit very large hysteresis loops accompanied by a corresponding conversion of energy to heat. Both these consequences have important effects on the properties of superconducting solenoids.

High-field superconductors in a physical state capable of carrying current differ in two respects from type I superconductors. Firstly, they do not exhibit a Meissner-Ochsenfeld effect. However, when they are initially exposed to a magnetic field, they can exclude much larger fields from the interior of the material, the amount being determined by the critical current density and size of the sample, than can type I superconductors. As the magnetic field strength is increased, the field progressively penetrates further and further into the sample. If the external field strength is reduced after substantial penetration, it is found that the magnetic field within the sample tends to remain trapped. This property of high-field superconductors has been considered in some detail by C. P. Bean [29]. Secondly, the current is not carried exclusively on the surface (figure 3) but more or less uniformly over the cross section of the material. The critical current should thus depend on the square of the diameter of a wire of the material, and the critical current density tends to be independent of the diameter, provided that the internal structure of the material remains unchanged.

High critical current densities are generally attained by introducing as many physical defects, such as dislocations, into the material as possible. Ductile materials, such as Mo-Re or Nb-Zr alloys, are drawn into wire, usually in the range 0·05–0·25 mm diameter, from a starting diameter as large as is practical (approximately 1–5 mm) without intermediate annealing. This type of operation is particularly effective in the creation of dislocations and related defects. The effect of mechanical deformation of Nb-Zr alloy containing 25 per cent

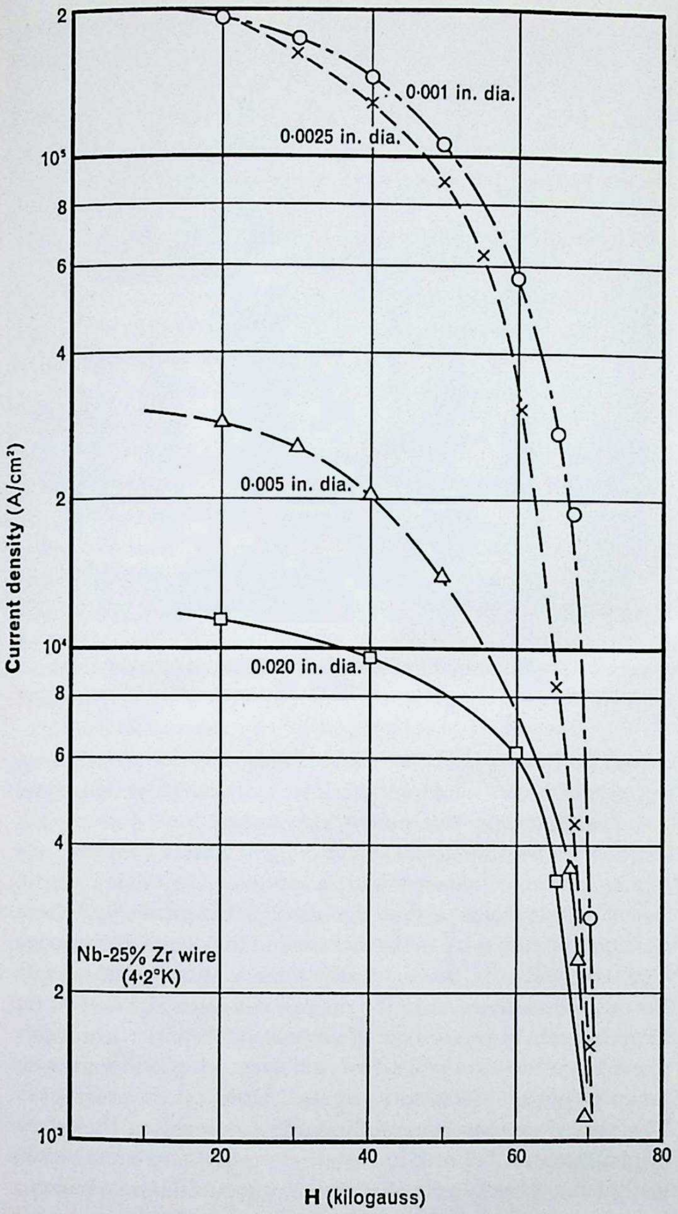


Figure 4 The pronounced effect on the critical-current density resulting from mechanical deformation (expressed in terms of wire diameter) of 3Nb-Zr [30].

Zr is illustrated in figure 4. The increase in critical current density by more than a factor of ten following the reduction in diameter to 0·006 cm is typical of the large effects observed during the mechanical reduction of ductile alloys.

The procedures for introducing a maximum number of dislocations and related defects in brittle materials such as Nb$_3$Sn are more subtle. In the case of Nb$_3$Sn, current densities in excess of 10^5 amp/cm^2 have been attained by carefully selecting the temperature at which the material is formed [31, 32]; particle size and time of reaction are also important factors. The process leading to the best results is believed to involve selection of a temperature (about 950°C) that represents a compromise between the rate at which Nb$_3$Sn is formed and the rate at which the dislocations and related defects are removed by annealing or recrystallization; both processes are favoured by increasing temperature. Critical current densities that have been achieved in this manner are higher, by a factor of more than fifty, than those of Nb$_3$Sn produced from a melt [1] involving temperatures above 2000°C.

Typical critical current densities at the boiling point of helium (4·2°K), as a function of magnetic field strength, are shown in figure 5 for several materials discussed above. In the case of Nb-Zr, the effect of deformation on critical current density is also apparent. However, the critical field in the limit of zero current density is not changed significantly by the mechanical deformation. This is reasonable, because the effect of deformation on the resistivity—and thus on the electron mean free path, l, and the effective coherence length ξ_{eff}—is usually small for disordered alloys whose resistivities are already high. If equation (3) is considered, it is apparent that a Nb-Zr alloy of the composition Nb$_3$Zr is not useful for producing a superconducting solenoid capable of creating fields over about 70 kG, because of the rapid decrease in critical current density at this field intensity. However, it could be used as the outer windings of a composite solenoid. The field at which the critical current density in Nb$_3$Sn decreases sufficiently for the material to be no longer useful for magnets is not accurately known. Steady magnetic fields of sufficient strength are not yet available to test the critical current density of Nb$_3$Sn beyond about 150 kG; from indirect evidence, it has been suggested that the critical field is well above 200 kG [7, 12, 33]. Pulsed field measurements by H. R. Hart and his co-workers have shown [34] that the critical current density in Nb$_3$Sn 'wires' remains sufficiently high to be useful for magnets up to field strengths of at least 190 kG.

Superconducting magnets

At present, the materials Nb-Zr and Nb$_3$Sn are being used extensively in superconducting solenoids. Each has advantages as well as limitations. Nb-Zr is ductile but is limited to fields of about 100 kG or less, though it should be possible to produce

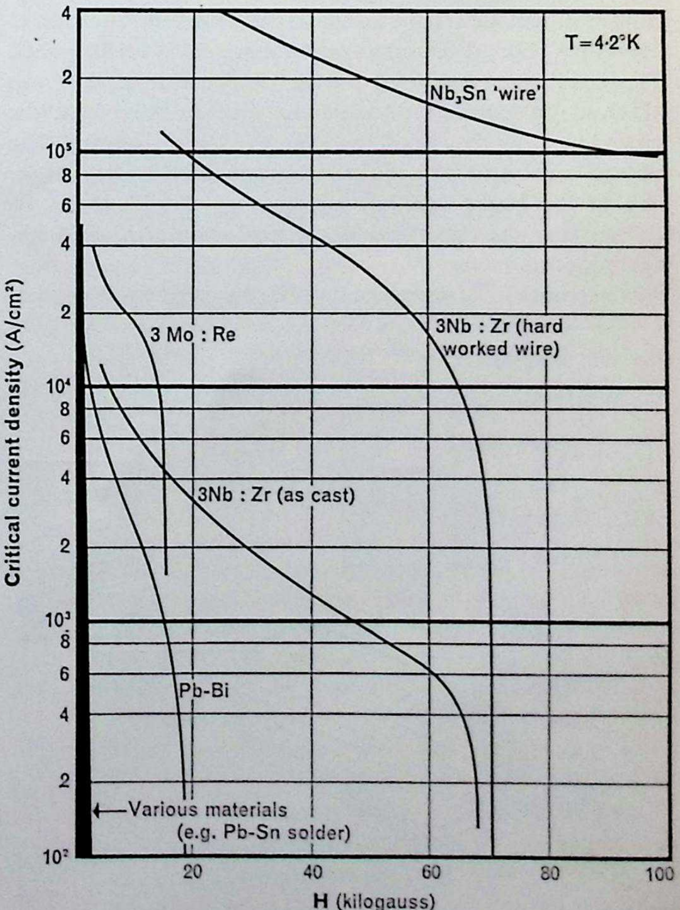


Figure 5 Typical critical-current density versus applied magnetic field characteristics, for several materials of interest for superconducting magnets.

fields above 70 kG by increasing the Zr content [12]. Nb-Zr magnets often show 'training' effects [35] and in practice sustain current densities that are significantly lower than those observed for small samples [35–37]. Nb_3Sn is brittle, and its fabrication into magnets presents major difficulties; however, the cored 'wire' process, described below, has been very successful. In addition, several other processes are under development, some of which do not require heating the finished magnet to near 1000°C, as does the cored 'wire' process. Magnets constructed from the cored 'wire' have not shown appreciable training effects nor deterioration of the solenoids when they are 'driven normal'.

Superconducting solenoids, using Nb-Zr for their windings, are being commercially produced by a number of companies, among them Avco, Linde, Magnion, Westinghouse, and Varian Associates. Fields as large as 70 kG can be obtained in working spaces having diameters of a few centimetres; working space diameters well over 10 cm are available at lower fields. The problem of Nb-Zr having a lower critical current density in coils than in short lengths has been partially overcome by copper-plating the wire. Using a heavy insulation, or spacing the windings further apart, also seems to be helpful. Although this 'proximity effect' is not well understood, it may be that the resulting decrease in the energy converted to heat per unit volume of windings, during the building up of the field, is a significant factor. In the case of the copper-clad wire, it is also conceivable that the existence of a low resistance shunt on the surface of the wire may locally lengthen the time constant. So doing would tend to minimize flux-jumping, which, because of the thermal energy produced, can lead to a chain reaction and consequently a premature 'driving of the coil normal'.

The right-hand side of figure 1 shows two early Nb-Zr coils studied at Bell Telephone Laboratories. Both have a throat of almost 0·5 cm and generate a maximum field of about 40 kG. The small coil was wound from 0·025 cm wire that was obtained by mechanically drawing 0·05 cm wire from the same batch as that used to wind the larger magnet. The critical current density of the smaller magnet is about twice that of the larger one (see equation 3) and illustrates, in another way, the importance of the mechanical state of a high-field superconductor.

The process used for preparing Nb_3Sn cored 'wire' is illustrated in figure 6. Finely divided Nb and Sn powders are intimately mixed and pressed into cylindrical pellets about 1 cm in diameter. The pellets are packed into a heavy niobium tube (wall thickness about 0·5 cm) about 5 metres long, which in turn is covered with a comparably thick jacket of monel metal; this is used as the outer jacket primarily because niobium tends to seize to the dies used in the drawing operation. The composite is mechanically drawn into 'wire' usually having a final diameter in the range 0·2–2 mm. The wire at this stage contains a ductile core of unreacted Nb and Sn; it is then wound into the form of a solenoid before being kept for several hours at about 950°C to form the brittle compound Nb_3Sn. The procedure now being followed is a refinement [38] of the original process [1] and has been extremely successful. Many continuous lengths of 'wire', each several thousand metres long, have been prepared. Although one would expect such a process to be susceptible to discontinuities in the core, little difficulty has been encountered. The few problems that have arisen have been found to be due to contamination by foreign material, improper mixing of the powders initially, and mechanical defects in the niobium and monel tubes that developed during drawing.

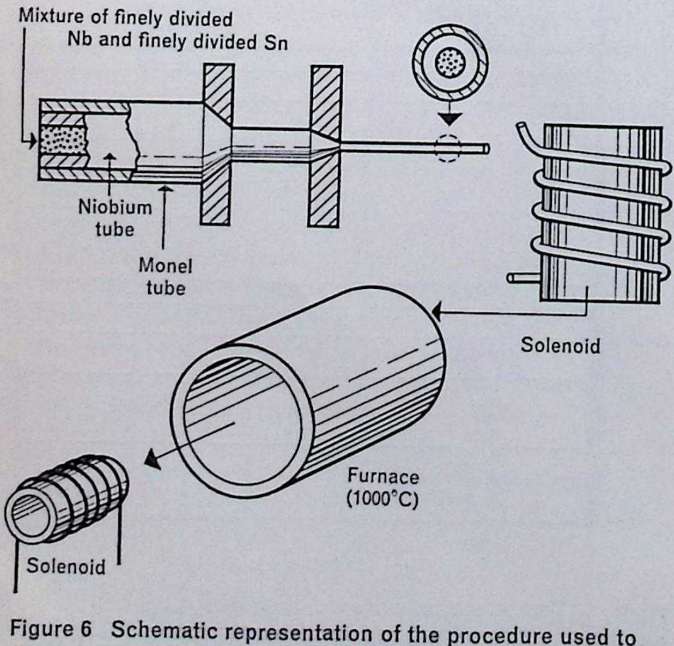


Figure 6 Schematic representation of the procedure used to fabricate Nb_3Sn-cored 'wire' and magnets.

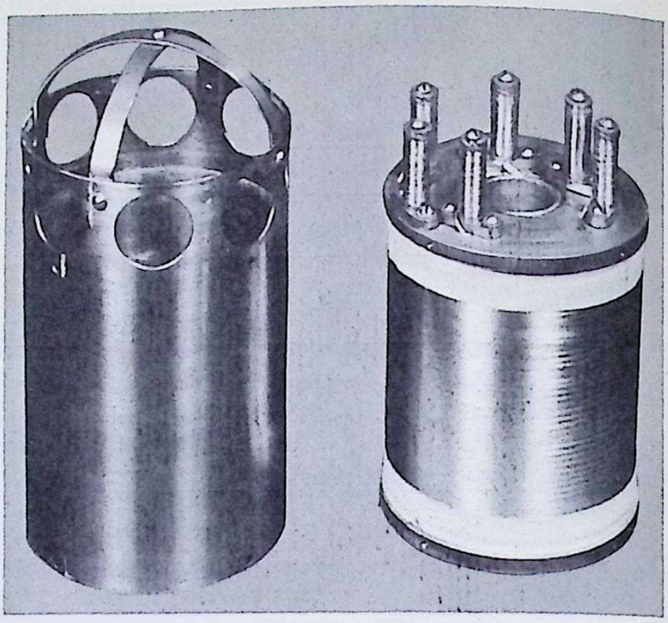


Figure 7 Nb_3Sn magnet with a 3 cm working diameter and capable of fields in excess of 70 kG; the magnet is 10 cm high.

On the left of figure 1 are shown three early Nb_3Sn magnets. All have throats of about 0·5 cm. The small one produces a maximum field of about 20 kG, the intermediate one about 40 kG, and the large one about 70 kG. Figure 7 shows a magnet (and its protective cover) having a throat nearly 3 cm in diameter. Magnets of this size provide useful working space for other investigations and are capable of fields of more than 70 kG.

Although there have been occasional difficulties in producing a coil with satisfactory properties, no Nb_3Sn-cored 'wire' magnets at Bell Telephone Laboratories have failed in testing or in use. Several dozen such magnets have been constructed and tested; some coils have been repeatedly 'driven normal' at least 15 or 20 times, and most have been cycled between room temperature and liquid helium temperature many times. This dependable behaviour is gratifying in view of the many potential sources of difficulty. The fact that difficulties have not so far developed is attributed to the favourable mechanical geometry of the Nb_3Sn-cored 'wire'. The brittle Nb_3Sn is constrained under pressure in the Nb jacket, because niobium is expected to have a larger coefficient of expansion; the composite is cooled from near 1000°C. The niobium jacket also provides a reasonably low resistance path—because it is a fairly pure metal and at a low temperature—for temporarily shunting the current past any weak spots in the winding during, and just preceding, the collapse of the magnetic field. The resistivity of Nb_3Sn in the normal state is relatively high. In addition, Nb_3Sn can withstand temperatures up to nearly 1000°C without deterioration, provided that thermal shock does not produce cracking. The Nb_3Sn-cored 'wire' coils have been wound using an insulation of low electrical resistance—an alloy of monel metal or some other low-resistance material—between the turns, even though the problem of the time constant has made it desirable to use a dielectric insulation, such as quartz-fibre cloth, between layers of windings. The time constant is $t = L/R$ (in seconds), where L is the inductance in henrys and R the resistance in ohms. This arrangement provides low-resistance paths for current flow during the collapse of the field when the coil is 'driven normal' and prevents the generation of excessive voltages. As magnets get larger in size and field strength, these problems become more acute, and additional difficulties must be anticipated.

Other techniques are being employed to fabricate Nb_3Sn into a form suitable for magnet construction. Investigators at the Radio Corporation of America have reported a method of depositing a thin layer of Nb_3Sn on wire substrates [39], and at the National Research Corporation tapes of Nb coated with Nb_3Sn are being produced. In both cases the Nb_3Sn layer is made thin, and the intention is to wind magnets not requiring subsequent heating. Models capable of fields as large as about 90 kG have recently been reported [40]. If these magnets prove to be practicable, there will be advantages in being able to avoid the problems associated with forming the Nb_3Sn by heating to near 1000°C after the magnets are wound.

In spite of the problems encountered, enough success has been achieved to remove any doubts concerning the feasibility of 100 kG superconducting magnets that are dependable and practical, and that have working volumes sufficiently large to be useful. It seems likely that these will be forthcoming shortly and that magnets capable of fields of near 200 kG are not very far off. It is very probable that the engineering problems associated with the generation of a large volume of a high field—for example, a cubic metre at 100 kG—will be formidable. Nevertheless, if a sufficiently compelling technological need is found to exist, the goals will doubtless be achieved.

High-field superconductors are too new for many of their potential applications to be apparent. Until the advent of these materials, it was generally not realistic for an engineer to think of designing devices requiring magnetic fields of 50–100 kG. Some of the more obvious potential applications include the generation of high magnetic fields for a wide variety of research activities, magnets for particle accelerators and bubble chambers, magnets for controlled thermonuclear fusion (should it prove to be feasible), magnets for magnetohydrodynamic power generation, high field-high current superconductors for high power d.c. transmission systems, and for the shielding and shaping of magnetic fields.

References

[1] Kunzler, J. E., Buehler, E., Hsu, F. S. L., and Wernick, J. H. *Phys. Rev. Letters*, 6, 89, 1961.
[2] Onnes, H. K. *Communs phys. Lab. Univ. Leiden*, **120b**, 1911; **122b**, 1911.
[3] *Idem. Ibid.*, **134b**, 1913; **133d**, 1913.
[4] *Idem. Ibid.*, **139f**, 1914.
[5] de Haas, W. J. and Voogd, J. *Ibid.*, **208b**, 1930; **214b**, 1931.
[6] Keesom, W. H. *Ibid.*, **234f**, 1935; *Physica*, 2, 35, 1935.
[7] Kunzler, J. E. *Rev. mod. Phys.*, 33, 1, 1961.
[8] Autler, S. H. *Rev. sci. Instrum.*, 31, 369, 1960.
[9] Shoenberg, D. 'Superconductivity'. (Second edition). p. 40. Cambridge University Press, Cambridge. 1952.
[10] Yntema, G. B. *Phys. Rev.*, 98, 1197, 1955.
[11] Kunzler, J. E., Buehler, E., Hsu, F. S. L., Matthias, B. T., and Wahl, C. J. *J. appl. Phys.*, 32, 325, 1961.
[12] Kunzler, J. E. *Bull. Amer. phys. Soc.*, 6, 298, 1961.
[13] Berlincourt, T. G., Hake, R. R., and Leslie, D. H. *Phys. Rev. Letters*, 6, 671, 1961.
[14] Wernick, J. H., Morin, F. J., Hsu, F. S. L., Maita, J. P., Dorsi, D., and Kunzler, J. E. in 'High Magnetic Fields'. (Edited by H. Kolm, B. Lax, F. Bitter, and R. Mills), p. 609. John Wiley & Sons, New York. 1961.
[15] Martin, D. L., Benz, M. G., Bruch, C. A., and Rosner, C. H. *Cryogenics*, 3, 114, 1963.
[16] Matthias, B. T., Geballe, T. H., Geller, S., and Corenzwit, E. *Phys. Rev.*, 95, 1435, 1954.
[17] Meissner, W. and Ochsenfeld, R. *Naturwissenschaften*, 21, 787, 1933.
[18] Bardeen, J., Cooper, L. N., and Schrieffer, J. R. *Phys. Rev.*, 108, 1175, 1957.
[19] London, F. 'Superfluids'. Vol. 1. John Wiley & Sons, New York. 1950.
[20] Mendelssohn, K. *Proc. roy. Soc., A*, 152, 34, 1935.
[21] Shaw, R. and Mapother, D. E. *Phys. Rev.*, 118, 1474, 1960.
[22] Hauser, J. J. and Buehler, E. *Ibid.*, 125, 142, 1962.
[23] Morin, F. J., Maita, J. P., Williams, H. J., Sherwood, R. C., Wernick, J. H., and Kunzler, J. E. *Phys. Rev. Letters*, 8, 275, 1962.
[24] Goodman, B. B. *I.B.M. J. Res. Develop.* 6, 63, 1962.
[25] *Idem. Phys. Letters*, 1, 215, 1962.
[26] Ginsburg, V. L. and Landau, L. D. Жур. эксп. теор. Физ., 20, 1064, 1950; Abrikosov, A. A. *Ibid.*, 32, 1442, 1957 [translation: *Soviet Phys., JETP*, 5, 1174, 1957]; Gor'kov, L. P. *Ibid.*, 37, 835, 1959 [translation: *Soviet Phys., JETP*, 10, 593, 1960].
[27] Kim, Y. B., Hempstead, C. F., and Strnad, A. R. *Phys. Rev. Letters*, 9, 306, 1962; *Phys. Rev.*, 131, 2486, 1963.
[28] Anderson, P. W. *Phys. Rev. Letters*, 9, 309. 1962.
[29] Bean, C. P. *Ibid.*, 8, 250, 1962.
[30] Olsen, K. M., Jack, R. F., Fuchs, E. O., and Hsu, F. S. L. in 'Superconductors'. (Edited by M. Tanenbaum and W. V. Wright), p. 123. Interscience Publishers, New York. 1962.
[31] Buehler, E., Wernick, J. H., Olsen, K. M., Hsu, F. S. L., and Kunzler, J. E. in 'Metallurgy of Advanced Electronic Materials'. (Edited by G. E. Brock), p. 105. Interscience Publishers, New York. 1963.
[32] Levinstein, H. J. and Buehler, E. (To be published.)
[33] Kunzler, J. E., Hsu, F. S. L., and Buehler, E. *Bull. Amer. phys. Soc.*, 6, 123, 1961.
[34] Hart, H. R., Jr., Jacobs, I. S., Kolbe, C. L., and Lawrence, P. E. in 'High Magnetic Fields'. (Edited by H. Kolm, B. Lax, F. Bitter, and R. Mills), p. 584. John Wiley & Sons, New York. 1961.
[35] Hulm, J. K., Fraser, M. J., Riemersma, H., Venturino, A. J., and Wien, R. E. in *Ibid.*, p. 332.
[36] Kunzler, J. E. in *Ibid.*, p. 574.
[37] Hake, R. R., Berlincourt, T. G., and Leslie, D. H. in *Ibid.*, p. 341.
[38] Olsen, K. M., Fuchs, E. O., and Jack, R. F. *J. Metals*, 6, 824, 1961.
[39] Hanak, J. J., Cody, G. D., Aron, P. R., and Hitchcock, H. C. in 'High Magnetic Fields'. (Edited by H. Kolm, B. Lax, F. Bitter, and R. Mills), p. 592. John Wiley & Sons, New York. 1961.
[40] Schroder, E. R., Freedman, N. S., and Fakan, J. C. *Appl. Phys. Letters*, 4, 105, 1964.

Molecular sieves

R. M. Barrer

It has been known for some time that zeolites can separate mixtures on the basis of the size and shape of the constituent molecules. However, it is only recently that the molecular structure that gives the zeolites this ability has been fully explored. This article discusses the molecular sieves, particularly in terms of the ways in which the sieving properties result from the structures of the compounds, and it suggests synthetic zeolites yet to be made that should have valuable properties.

The capacity of porous solids to act as sieves on a molecular scale, that is, to sort mixtures of molecules according to their size and shape, has been exhibited in a variety of sorbents. Thus partial molecular sieve sorption is found with certain coals [1], specially prepared active carbons [2], and silica gels [3]. Nowhere, however, has molecular sieve separation been found to be so specific and quantitative as in some of the naturally occurring aluminosilicates known as zeolites, and in synthetic varieties of these.

As subjects of study, zeolites have long attracted mineralogists and geologists, and early work on sorption by zeolites has been summarized by J. W. McBain [4]. In the late thirties [5] their sorbent properties began to be systematically examined by the author, and by the middle forties a thorough-going demonstration had been provided of the unique ability of such zeolites as chabazite and mordenite to separate mixtures [6]. The examples given in this reference provide illustrations of many of the uses of today, such as the purification of organic chemicals, and separation of straight-chain hydrocarbons from branched-chain, aromatic, and naphthenic hydrocarbons. There is a general similarity of technique in all these applications. The mixture is passed through a vessel that contains the zeolite, and emerges enriched in one component. The other is subsequently removed from the zeolite, if required, by changing the pressure or temperature in the vessel, or by displacement with another sorbate. Synthetic zeolites were made (analcite, mordenite, and a synthetic zeolite without natural counterpart), and this work has been continued to the present time, yielding a rich variety of species [7].

It was shown in 1947 that molecular sieving at low temperatures was sufficiently sensitive to separate argon from oxygen [8], and this led to work in the United States that has culminated in the large-scale production and marketing of a range of Linde molecular sieves; this range has been augmented by the Zeolons of the Norton Company. Work by G. T. Kerr and G. T. Kokotailo [9] of Socony Oil, and also by the author and his colleagues [10], has yielded other synthetic zeolites with some novel properties. Research on the new sorbents has started in Germany, France, Czechoslovakia, and Japan, and the Soviet Union has established a Zeolite Commission, which has launched a many-sided programme for research and development. Meanwhile, work in the author's laboratory revealed how one might further 'tailor' a given molecular sieve sorbent. There are several possible methods which are described later [11].

Thus a new and many-sided attack on zeolite problems has developed. Many of the properties of zeolites depend on their structure, and growth of interest in their chemistry has stimulated investigations that have revealed much of their remarkable molecular architecture [12–16]. Their ion-exchange behaviour has been examined from the viewpoints of equilibrium, kinetics, and selectivity [18, 19]. In certain cases they have shown that they can scavenge cations from radioactive effluents and hence remove dangerously radioactive constituents. Lately geologists have reported massive zeolite deposits [20] which may compete in certain respects with the synthetic materials. Because of their comparative abundance in basic rocks and in certain sediments, it has recently been suggested that zeolites can play a significant part in geological processes of mineral growth and transformation [21].

R. M. Barrer, *D.Sc., Sc.D., F.R.S.*
Was born in 1910 and educated at Canterbury College, New Zealand, and the University of Cambridge. After a period as a Research Fellow at Clare College he became successively Head of the Chemistry Department at the then Technical College in Bradford; Reader in the University of London; Professor of Chemistry, Aberdeen University; and since 1954 Professor of Physical Chemistry and Head of the Department of Chemistry at Imperial College, London.

Naturally occurring zeolites

There are over thirty naturally occurring zeolites. They are found in basaltic rocks rich in alkalis, such as occur in Northern Ireland and Iceland [22]. The zeolites have been formed in such media by hydrothermal alteration, sometimes as beautiful crystals lining small cavities (figure 1). They are also found in massive amounts as alteration products in sediments [20]. The zeolites in the late Cenozoic sediments in the United States are products of low-temperature reactions between a sediment and saline lake waters (phillipsite, erionite, clinoptilolite, and analcime). The triassic sediments from Southland, New Zealand, contain widespread beds rich in laumontite, heulandite, or analcime, and show that zeolitization can be an important stage in progressive regional metamorphism.

There are also purely synthetic zeolites, and synthetic variants of natural species have been made in which chemical compositions, optical properties, and, most importantly, the sorptive properties may differ from those of their naturally occurring counterparts. However, all zeolites have readily exchangeable cations, and intracrystalline water which is lost, continuously and often reversibly, as the temperature is raised. They may be given the empirical oxide formula $(\mathrm{M^{II}}, \mathrm{M_2^{I}})\mathrm{O.Al_2O_3}.n\mathrm{SiO_2}.m\mathrm{H_2O}$. The values of n and m may vary: $\mathrm{M^{II}}$ and $\mathrm{M^{I}}$ are respectively di- and monovalent cations (Ca, Sr, Ba, Na, K). The ratio $(\mathrm{M^{II}}, \mathrm{M_2^{I}})\mathrm{O} : \mathrm{Al_2O_3}$ is unity, which means that the structures are all based on $\mathrm{AlO_4}$ or $\mathrm{SiO_4}$ tetrahedra linked to each of four other tetrahedra by sharing the four apical oxygen atoms. According to an empirical rule of W. Loewenstein [23], $\mathrm{AlO_4}$ tetrahedra can be joined in this way to $\mathrm{SiO_4}$ tetrahedra only, never to another $\mathrm{AlO_4}$. This means that n in the oxide formula cannot be less than two in the completed aluminosilicate framework. In special circumstances, discrete $\mathrm{NaAlO_2}$ molecules can be trapped in some of the interstices of the framework and may cause apparent anomalies of composition. However, these molecules are not part of the framework [24]. The cations $\mathrm{M^{II}}$ and $\mathrm{M^{I}}$ and the water molecules are also present in interstices of the framework. This carries a net negative charge equal to the number of Al atoms in it.

In directly synthesized zeolites, the ions $\mathrm{M^{II}}$ and $\mathrm{M^{I}}$ may, in addition to those already mentioned, be $\mathrm{Rb^+}$, $\mathrm{Cs^+}$, or $\mathrm{NH_4^+}$, or simple organic ions such as $(\mathrm{CH_3})\overset{+}{\mathrm{N}}\mathrm{H_3}$, $(\mathrm{CH_3})_2\overset{+}{\mathrm{N}}\mathrm{H_2}$,

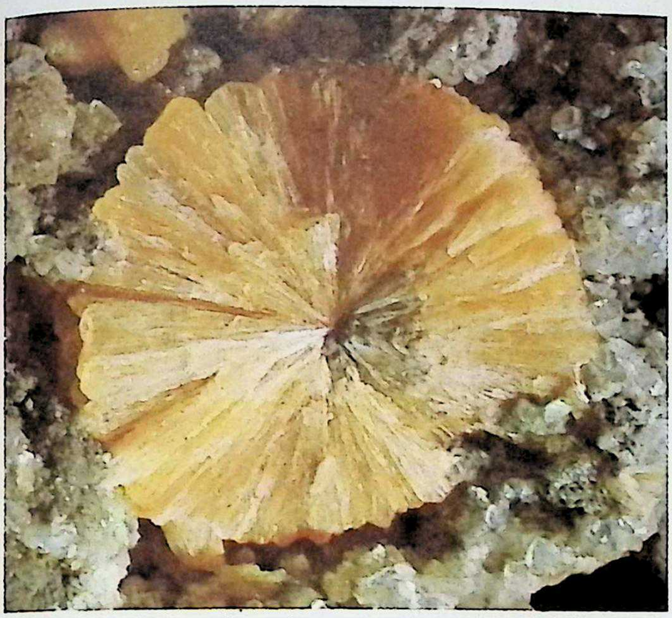


Figure 1a Stilbite (yellow sheaves), with some heulandite.

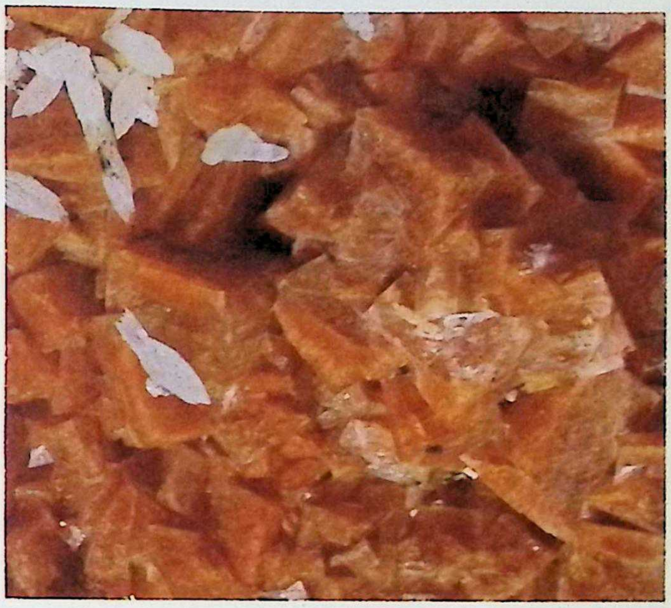
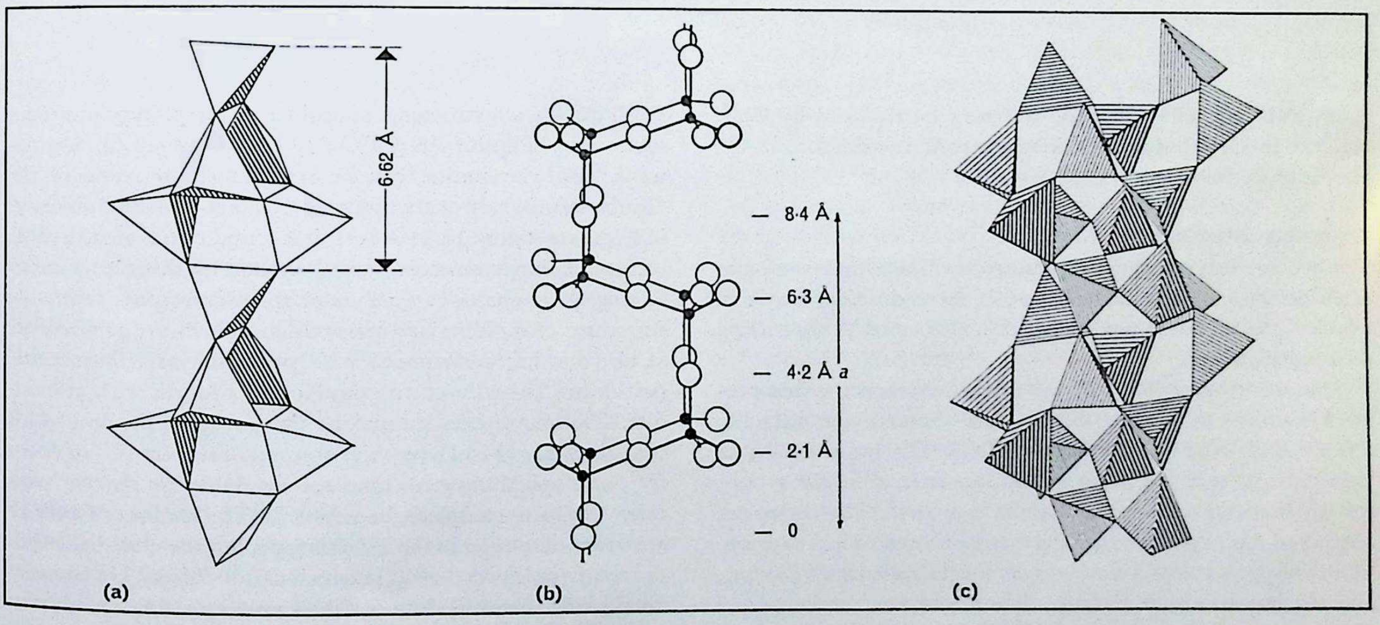


Figure 1b Acadialite, a form of chabazite. (By courtesy of the Trustees of the British Museum (Natural History).)


6·62 Å
8·4 Å
6·3 Å
4·2 Å a
2·1 Å
0
(a)
(b)
(c)

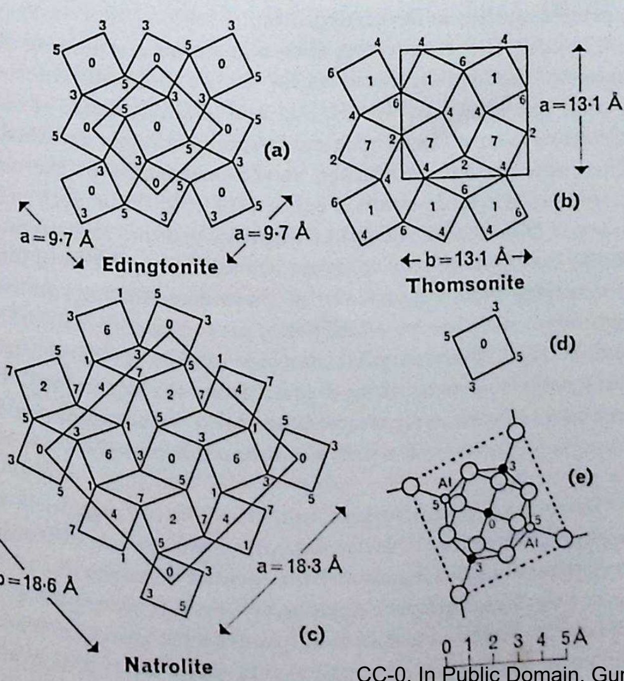
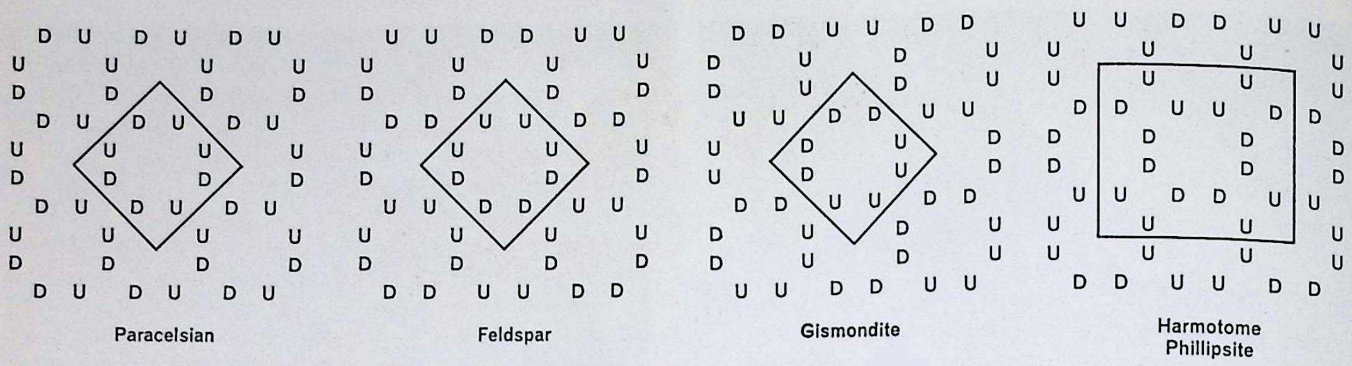

(a)
a=9·7 Å
a=9·7 Å
Edingtonite
a=13·1 Å
(b)
b=13·1 Å
Thomsonite
(d)
(e)
Al
Al
b=18·6 Å
a=18·3 Å
(c)
Natrolite
0 1 2 3 4 5Å

Figure 2 Above: Three kinds of chains of linked tetrahedra found in different zeolite groups: (*a*) the natrolite group chain, (*b*) the phillipsite group chain, (*c*) the mordenite group chain.

Figure 3 Left: The cross-linking of the chains for several members of the natrolite group. (After Bragg, 'Atomic Structure of Minerals', Cornell University Press, 1937.)



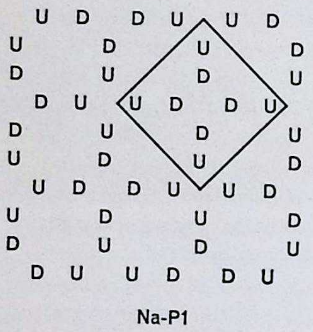


Figure 4 A formal representation of the cross-linking of 4-rings in different phillipsite chains in some minerals based on the crankshaft chains of figure 2*b* (paracelsian to gismondite). The chains run normal to the plane of the paper. The *a* and *b* dimensions of the unit cells are delineated as squares, and different types of 8-rings are seen in the figure [27]. The letters *D* and *U* represent tetrahedra with apices pointing respectively down and up. Na-P1, though based on 4-rings, does not contain crankshaft chains.

$(CH_3)_3\overset{+}{N}H$, or $(CH_3)_4\overset{+}{N}$. Also Al^{III} may be replaced by Ga^{III}, and Si^{IV} by Ge^{IV}, in the anionic framework. In addition, H_3O^+ may replace Na^+ by ion exchange.

Crystal architecture

The geometrical patterns of the aluminosilicate framework can often be visualized in at least one of three different ways: in terms of chains, layers, and polyhedra, all formed by the linking of tetrahedra.

Structures representable in terms of chains. Among the first zeolite structures to be established were those of the natrolite ($Na_2O.Al_2O_3.3SiO_2.2H_2O$) group [25]. The same chains of tetrahedra, illustrated in figure 2*a*, appear in all of the group; the figure shows two repeating units in a chain. The chains are combined laterally with one another by shared apical oxygens. The chain unit comprises a group of five tetrahedra (figure 3*e*), and if one views such a chain end-on, one can see four-sided rings (figure 3*d*) transverse to the chain direction (which is along the *c*-axis). Representing these 4-rings as squares, the structures of edingtonite ($BaO.Al_2O_3.3SiO_2.3H_2O$), thomsonite ($(Na_2Ca)O.Al_2O_3.2SiO_2.3H_2O$), and natrolite are, as illustrations, shown in figure 3. The numbers in the corners of the squares give the heights of the appropriate Al and Si atoms as multiples of $c/8$, where c is the repeat distance along the *c*-axis, 6·6 Å. A comparison with figures 3*d* and 3*e* shows that the corners of the squares represent oxygen atoms, and that the silicon and aluminium atoms referred to lie, at their various levels, inside the squares. The chains enclose, in this direction, rather narrow channels, which in hydrated natrolite have a minimum free diameter of about 2·1 Å. The openings between neighbouring chains in natrolite are somewhat larger ($\sim$2·6 Å minimum free diameter), resulting in an intersecting net of channels [26]. The density of the bonds tying the chains together is less than that of the bonds in individual chains, so that the natrolite family is somewhat fibrous in character. The chains can rearrange themselves during dehydration, but when fully hydrated, the free volume available to water molecules ranges between extremes of 0·21 (natrolite) to 0·33 (thomsonite) cm^3 per cm^3 of crystal.

All the known structures of zeolites of the phillipsite group, such as phillipsite ($(K_2Ca)O.Al_2O_3.4SiO_2.4{\cdot}5H_2O$, harmotome, and gismondite, can be represented in terms of the 'double-crankshaft' chain (figure 2*b*). This chain is also observed in feldspars (figure 4). However, it is found that a considerable variety of simple structures can be made by differently cross-linking these chains [27]. Two of them correspond with the structures of sanidine and paracelsian, which are not zeolites at all; another corresponds with phillipsite and harmotome (which are virtually isostructural); and a fourth with gismondite. The four tetrahedra making the 4-rings of the crankshaft chain in figure 2*b* can have their apices pointing up (*U*) or down (*D*), and apical oxygen atoms are available for sharing with other chains to complete the networks. The chains of figure 2*b* are based on 4-rings of the *UUDD* type, but the cross-linking of the chains produces 8-rings in which various U and D sequences are possible (figure 4). Many of the seventeen possible structures represent species so far unidentified.

The fifth of the structures shown in figure 4 represents that suggested for the synthetic zeolite Na-P1 [28], in which the 4-rings are alternatively *UUUU* and *DDDD*. Pairs of rings accordingly combine to give cubic groups of eight tetrahedra. The cubic groups are located at the centre and at the eight corners of the body-centred cubic unit cell. Thus each cubic group is joined to one of eight other such groups. This produces the 8-rings that are seen in figure 4, and also a system of interpenetrating channels running in three directions mutually at right angles. Other possible 4-ring arrangements are *UDUD* and *UUUD*. Four simple structures can be developed from the former ring pattern, and seven from the latter [27]. The treatment of various structures based on 4-rings is an interesting example of the way in which hitherto unidentified species may be predicted.

There may be distortions from the ideal structure in the network. For example, in the anhydrous frameworks of sanidine and paracelsian, the chains have twisted so much that little intracrystalline free space remains. In harmotome or phillipsite, the chains have twisted so as to produce the opposite effect—an even more open channel system than that in the ideal

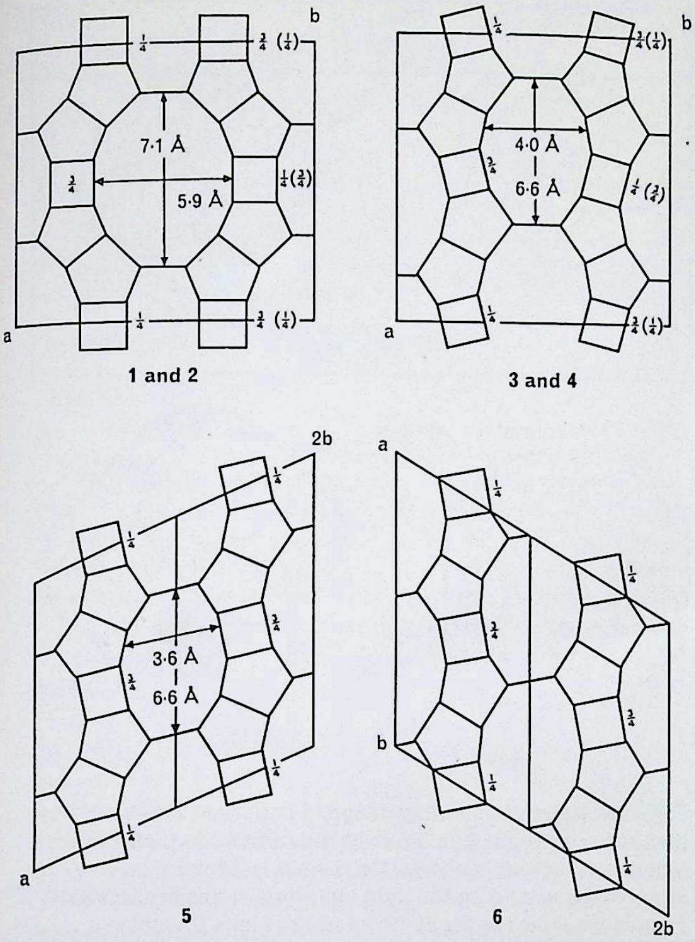


Figure 5 Six simple ways of linking different mordenite chains of figure 2*c* [14]. The chains run in the *c*-direction, normal to the plane of the diagram, and as shown Al or Si atoms are centred at each apex. Free dimensions are given for some of the channels which run parallel to the *c*-axis. The fractions denote the heights of the planes of the 4-rings in units of *c*, measured along the *c*-parameter of the unit cell.

framework. Zeolites of the phillipsite group have intracrystalline volumes for water ranging between 0·34 (harmotome) and 0·49 (gismondite) cm^3 per cm^3 of crystal.

The third type of chain, found in the mordenite group [14], is illustrated in figure 2*c* in terms of the individual tetrahedra, SiO_4 or AlO_4, of which they are composed. There are six possible simple structures formed by different lateral bondings of the chains to one another, and cross-sections of the resultant unit cells are shown in figure 5 [14]. All contain particularly wide non-intersecting channels running parallel to the *c*-axis, which is the direction normal to the plane of the figure. Structure 1 corresponds with mordenite, the channels in which are seen to be circumscribed by 12-rings. Of the other modes of cross-linking the chains, that of number 5 on this figure corresponds with the rare zeolite dachiardite [14]. The remaining four structures have not so far been identified in any zeolite, and again represent a challenge for the chemist. Two of the structures have channels circumscribed by 12-rings and four by 10-rings. Other cavities may also exist in these crystals. For example, each of the main channels in mordenite is lined with a double row of pockets opening into it, each big enough to accommodate molecules of permanent and noble gases. Mordenite and dachiardite have free volumes accessible to water molecules of about 0·3 cm^3 per cm^3 of crystal.

Structures representable in terms of layers. Two felspathoids, sodalite and cancrinite, have open aluminosilicate frameworks which in the naturally occurring specimens are filled with salts, such as $NaCl$, Na_2SO_4, $CaSO_4$, and $CaCO_3$. However, one can synthesize basic sodalites and cancrinites under hydrothermal conditions, in which these salts are replaced by water molecules and a little caustic soda [29]. The products then behave as zeolites.

Cancrinite and sodalite have aluminosilicate frameworks representable in terms of sheets of linked rings of six tetrahedra, shown in idealized form in figure 6 [30]. Each of these layers is firmly united by means of Si—O—Si or Al—O—Si bonds with identical sheets above and below. In cancrinite the sequence of the layers is 1212 . . ., that is, each third layer is superposable on the first, while in sodalite the sequence is 123123 . . ., each fourth layer being superposable on the first. The result in cancrinite is to create a structure having wide channels with a free diameter of about 6 Å, all running parallel to the *c*-axis and circumscribed by 12-membered rings. In sodalite, the arrangement produces a structure with cavities having free diameters of 6·6 Å. These cavities have the shape of a truncated octahedron (figure 8*a*), and indeed sodalite can be equally well represented by a simple packing of these figures, which have been called cubo-octahedra [31].

In chabazite and gmelinite one may recognize aluminosilicate sheets containing hexagonal prism units in which two 6-rings are united by shared apical oxygens to give the prism [15]. Each layer of prisms is very firmly joined to identical layers above and below, by pairs of bonds (Al,Si)—O—Si forming 4-rings. A sequence of layers that repeats itself in alternate layers is called a 1212 . . . sequence, and this occurs in gmelinite, giving a structure which, like cancrinite, has wide channels running parallel to the *c*-axis, circumscribed by 12-rings and having free diameters of about 6·6 Å. However, no one has yet obtained the molecular sieve character predictable for such channels in either basic cancrinite or in gmelinite [15, 32]. It may be that there are stacking faults, in which the 1212 . . . sequence of layers is broken by an occasional 123 sequence, which interrupt these channels. Another reason may be that cations and (in basic cancrinite) caustic soda are

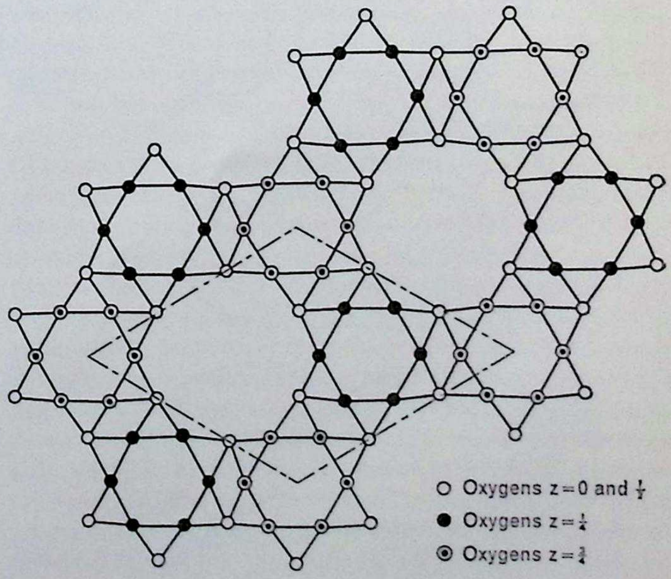


Figure 6 The hexagonal network characteristic of the sodalite-cancrinite minerals [30]. 6-rings are linked through 4-rings to give the puckered sheets of the diagram, enclosing wide tunnels running parallel to the *c*-axis, which is normal to the plane of the paper. The heights of the oxygen atoms are expressed as fractions of the *c*-parameter of cancrinite, which is ~ 5·2 Å.

present in the wide channels, and partially block them. The layer sequence in chabazite is 123123 . . . and results, as in the case of sodalite, in a structure containing substantial cavities, one of which is shown in figure 7*a* [33]. These cavities are no longer approximately spherical as in sodalite, but are elongated, and have free dimensions of 6·6 Å and ~ 11 Å.

There are other structures in which layers of single hexagons, as in sodalite-cancrinite minerals, alternate with layers containing hexagonal prisms as in chabazite and gmelinite. Such are erionite [16], in which the sequence[1] is 1′21′41′21′4 . . ., and levynite, in which it is 1′23′45′61′23′45′6 . . . [33]. This produces main cavities in erionite with free dimensions of about 6·3 Å and 15·1 Å as shown in figure 7. The cavities in levynite are roughly spheroidal, with the maximum free diameter about 7·8 Å (figure 7).

Subsidiary cavities are formed both in erionite and in gmelinite. In the latter these are large, with free dimensions of about 7·8 Å and 6·5 Å (figure 7). In erionite they are quite small.

It is possible to predict structures [27], not so far reported, which are based upon either one or both of the types of layer considered in this section. Thus, if we consider only the sodalite-cancrinite 6-ring layers, two possibilities are 1213 . . . and 12123 . . . The second of these is a mixed cancrinite-sodalite phase in which bands of each mineral alternate. Similarly, based only on the hexagonal prism layers in chabazite, we can have 1′2′1′3′ . . . and 1′2′1′2′3′ . . ., the last being a banded chabazite-gmelinite phase. Further structures containing layers of both kinds include the sequences 1′2 . . ., 1′23 . . ., 1′2′3 . . ., 1′213 . . ., 1′23′2 . . ., 1′213′ . . ., 1′232′ . . ., 1′2123 . . ., and like arrangements. It is not difficult to see the kinds of cavity which these hypothetical structures will contain.

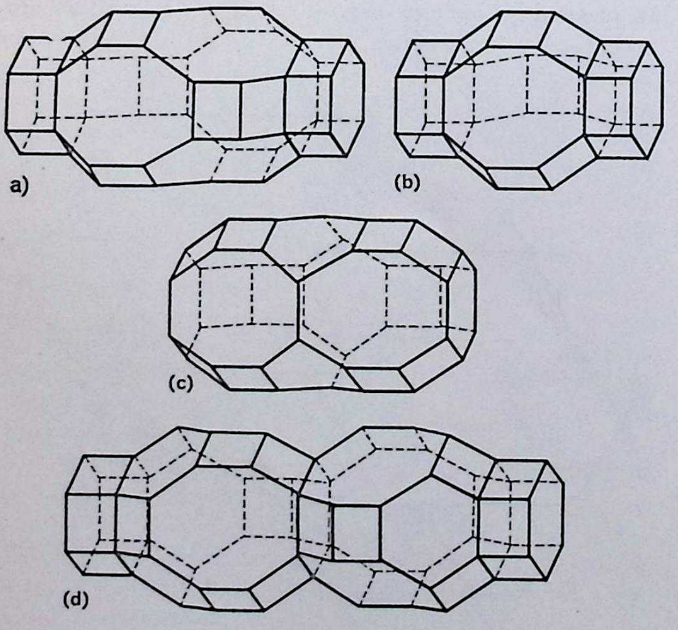


Figure 7 Cavities in several zeolites of the chabazite group [33]: (*a*) chabazite, (*b*) gmelinite, (*c*) erionite, (*d*) levynite. Al or Si atoms are centred on each apex, and oxygens which are not shown are centred near but not on the midpoints of all edges. In gmelinite, in addition to the cavity shown, there are wide channels running parallel to the *c*-axis, while in erionite there are subsidiary smaller cavities.

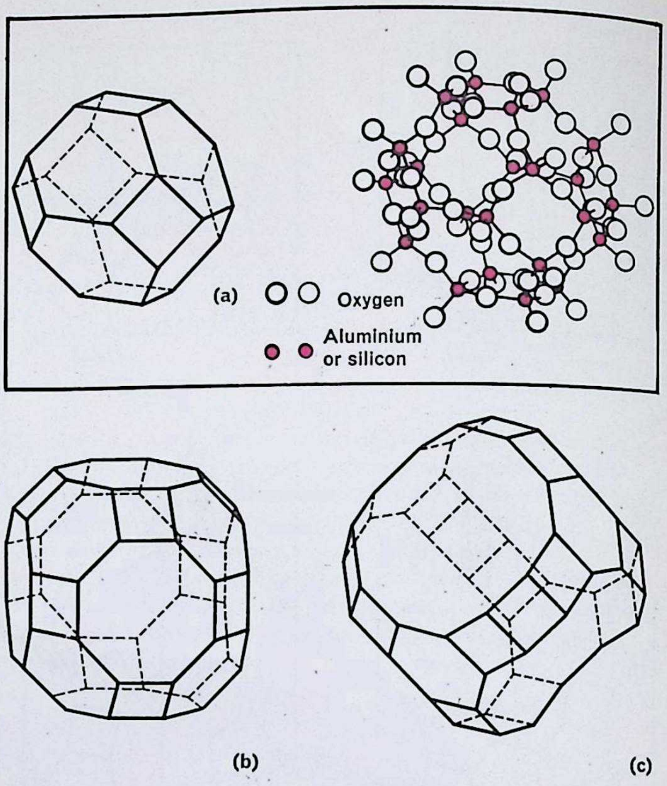


Figure 8 Polyhedral cavities found in important zeolites other than those in figure 7. In figure 8*a* the truncated octahedron (cubo-octahedron) is shown. On the left is a formal representation, and on the right individual atoms are indicated. Figures 8*b* and *c* are the 26-hedra of two kinds found in zeolite A and ZK-5, and in faujasite and zeolites X and Y respectively. Same scale as figure 7.

Structures which can be formed from polyhedra. Various zeolites such as faujasite, Sieve A, and zeolites ZK-5, Na-P 1, and basic sodalite are conveniently envisaged in terms of secondary building units in the form of polyhedra, appropriately linked and stacked. A number of these, and examples of structures in which they occur, are given in table 1. The relation of the cubo-octahedron to the individual tetrahedra composing it is shown in figure 8*a* [34]. In the line diagram on the left the corners are occupied by Al or Si atoms, and the oxygens are not shown. The same convention is adopted in figure 7, and in figure 8*b*, which shows the truncated cubo-octahedron and the 26-hedron [35]. These larger polyhedra have very substantial free space, and the diameters of some of the inscribed spheres are also shown in table 1. Clearly there is ample room for guest molecules in these cavities and in the cavities shown in figure 7. Other polyhedra, such as the octagonal prism, are possible, as also is a polyhedral cavity bounded by twelve 4-rings and six 8-rings (zeolite ZK-5 [36]).

A simple way to visualize the sodalite framework is to stack the cubo-octahedra of figure 8*a* in 8-fold co-ordination, so that a given central cubo-octahedron shares one of its eight 6-ring faces with one each of eight other cubo-octahedra. In the same way, Sieve A can be visualized as composed of truncated cubo-octahedra (figure 8*b*) stacked in 6-fold co-ordination, a central polyhedron sharing one of its six 8-ring faces with one each of six other such polyhedra. The faujasite structure can be made by stacking the 26-hedra of figure 8*b* in 4-fold co-ordination, a central 26-hedron sharing one of its four 12-ring faces with one each of four other 26-hedra.

The cubo-octahedron stacked as described in the previous paragraph fills all space. The packing of truncated cubo-octahedra described above for Sieve A leads to other polyhedra. These are cubo-octahedra and cubic units, which, together

---

[1] The layers marked thus, ′, contain hexagonal prisms.

Table I
Some polyhedral units present in zeolites

Unit	Number of SiO_4 and AlO_4 tetrahedra in unit	Number and type of faces		Free diameter of inscribed sphere (Å)	Examples of occurrence
		No.	Type		
Cube	8	8	4-rings	—	Zeolite Na-P1 and Sieve A
Hexagonal prisms	12	6 2	4-rings 6-rings	—	Chabazite, gmelinite, levynite, erionite, and faujasite
Cubo-octahedron[1]	24	6 8	4-rings 6-rings	6·6	Sodalite, Sieve A, faujasite
Truncated cubo-octahedron (26-hedron, type 1)	48	12 8 6	4 rings 6-rings 8-rings	11·4	Sieve A, Zeolite ZK-5
26-hedron, type 2	48	18 4 4	4-rings 6-rings 12-rings	11·8	Faujasite

[1] Also referred to as a truncated octahedron.

with the truncated cubo-octahedra, then fill all space. Indeed, the Sieve A structure can be equally well described in terms of cubo-octahedra in 6-fold co-ordination with one another, and joined through their 4-ring windows by four Si—O—(Si,Al) bridges. This produces cubic units as the bridges between cubo-octahedra [31]. Similarly, the stacking of 26-hedra in faujasite as indicated in the previous paragraph also produces cubo-octahedra and hexagonal prisms, and these three polyhedra together fill all space. The faujasite structure can therefore equally well be described in terms of linked cubo-octahedra. Four cubo-octahedra are linked to any central cubo-octahedron by six Si—O—(Si,Al) bridges utilizing four out of the eight 6-ring faces of the central cubo-octahedron. Thus each cubo-octahedron is located as is the carbon atom in diamond, and the bonds between cubo-octahedra are hexagonal prisms. The 26-hedra play a role in the structure like that of the interstices between the carbon atoms in diamond.

Zeolite ZK-5 is readily described [36] in terms of the truncated cubo-octahedron of figure 8*b*. A central polyhedron is linked through its eight 6-ring faces with each of eight other polyhedra by six Si—O—(Si,Al) bridges. Thus the bonds uniting pairs of truncated cubo-octahedra are hexagonal prisms, while the arrangement produces additional polyhedral cavities bounded, as mentioned previously, by twelve 4-rings and six 8-rings. It is interesting that a central truncated cubo-octahedron can also be linked by octagonal prism 'bonds' to each of six other truncated cubo-octahedra to form a new structure so far not synthesized [36]. By recognizing the building units characteristic of some known species and by fitting these together in different ways, one may anticipate possible chemical discovery and visualize novel kinds of crystal architecture suitable for molecular and ion sieving and other purposes.

Faujasite and Sieves A and ZK-5 are highly porous, having free volumes accessible to water of about 0·53, 0·46, and 0·41 cm³ per cm³ of crystal. In this they resemble chabazite and gmelinite (~ 0·46–0·45 cm³ per cm³).

Guest molecules

In view of what has been said about the porous structures of zeolites it is hardly surprising that, after removal of zeolitic water by heating and outgassing, they are in many cases able to imbibe guest molecules in great variety and yet in a highly selective manner. The crystals are permeated by continuous channel systems which are often interpenetrating, and entry into these is regulated by certain windows located periodically through the structures. These windows often separate larger cavities which are themselves part of the channel systems. In table II are given the approximate numbers of some guest molecules, the sorption of which has actually been studied, which fill the larger cavities in typical molecular sieve zeolites. Several properties of interest can be seen from this table.

Firstly, as noted above, the variety and dimensions of the sorbates listed in the last column are controlled by the free dimensions of the windows giving access into the cavities and not by the free dimensions of the cavities themselves (see table I). Thus, basic sodalite sorbs only small polar molecules such as water or ammonia, and sorbs these quite sluggishly. In this substance, an energy of activation is involved in passage through the 6-ring windows even for small molecules. Chabazite presents non-circular apertures to potential sorbates, and although it behaves rather like Sieve A, its permeation by many species is considerably more sluggish. On the other hand, the remarkably wide apertures giving access to the cavities in faujasite allow even large molecules to permeate the crystals freely.

Secondly, it may be observed from the table that clusters of molecules, frequently of considerable size, are present in the cavities when these are saturated. Moreover, these clusters may be joined, however tenuously, by contacts with other clusters through the windows. Such contacts must be considerable where, as in faujasite, wide apertures link one large cavity to each of four others, and one may have not just partially isolated clusters, but continuous filaments of zeolitic fluids [38]. The number of molecules in any cluster is not an integer, because a molecule may be shared between two cages if it happens to be sited in the windows joining two cavities.

When the cavities are not saturated, and the numbers of guest molecules are small, they are distributed, not necessarily uniformly, throughout the whole accessible pore volume. The sorbate behaves like a gas in that it fills the pore volume available to it at a mean density which increases as the amount of sorbate increases. When the intracrystalline pore volume is nearly saturated with guest molecules, the packing density of these molecules per unit volume approaches that in the bulk liquid [39]. Some variations may arise because of steric difficulties in the arrangement of the larger guest molecules in a spatially economical way in pore spaces not much larger than themselves. The coefficient of thermal expansion at constant sorption potential $\Delta\mu$ (defined as equal to $RT \ln p/p_0$) also approaches that of the bulk liquid (p is the equilibrium pressure of guest species over the host crystals, and p_0 the saturation vapour pressure of liquid guest species). This means that the apparent saturation numbers of guest molecules varies with temperature approximately as the density of the bulk liquid, and is a reason why saturation numbers in table II are to be considered as approximate only. They mostly refer to temperatures not far removed from room temperature. In the faujasite-like compound Na-X, the apparent saturation numbers for ammonia in one-eighth of a unit cell, that is, in a 26-hedron (table I) plus a cubo-octahedral cavity, were found to vary as follows [40]:

Temperature (°C)	0	30	50	70	100	120	140	160	180
Saturation number of NH_3 molecules	23·2	21·3	19·6	18·3	16·9	16·0	14·6	13·3	11·7

Where the thermochemistry of inclusion of guest molecules has been investigated, their binding energy within the crystals has been shown to depend, for non-polar species such as the noble gases, upon dispersion, close-range repulsion, and polarization energies [35, 38–41]. If the guest molecules have permanent electric moments, then field-dipole and field gradient-quadrupole terms become significant. The anionic oxygens of the aluminosilicate framework, being rather polarizable and present in the greatest numbers of all atomic constituents of the host crystals, dominate the dispersion energy, but the cations in the zeolite greatly influence polarization, dipole, and quadrupole energies. It is found that the heat of inclusion then varies strongly with amount sorbed, so that zeolites are often energetically inhomogeneous as sorbents. The heteropolar environment within the crystals may activate the sorbates and effect catalytic transformations (oxidation, cracking, polymerization, or disproportionation) [42]. Not only does the crystal influence the guest species, but the converse behaviour can also be observed. The electrical conductivity and dielectric loss of zeolites depend strongly upon their state of hydration; indeed the conductivity may change even in the presence of such inert sorbates as nitrogen or oxygen [43].

Table II

Saturation numbers of guest molecules in typical polyhedral cavities

Crystal and cavity	Windows controlling access to cavity	Approximate free dimensions of windows (Å)	Approximate number of guest molecules per cavity
Basic sodalite (cubo-octahedron)	6-rings	2·2	4 H_2O+0·5 NaOH
Chabazite (20-hedron, with six 8-ring, eight 6-ring and twelve 4-ring faces)	8-rings (non-polar)	3·1 (min.) and 4·4 (max.)[1]	12–14 H_2O 7·7 NH_3 6–7 Ar, N_2, O_2 4·9 CH_3NH_2 4·3 CH_3Cl 3·1 CH_2Cl_2 2·0 I_2
Sieve A (truncated cubo-octahedron, table I)	8-rings (planar)	4·3	29 H_2O (25+4)[2] 19–20 NH_3 14–16 Ar, N_2, O_2 15 H_2S 12 CH_3OH 10 SO_2 9 CO_2 5·5 I_2 5·4 n-C_3H_7OH 4 n-C_4H_{10}
Faujasite and Sieves X and Y (26-hedra, table I)	12-rings (near-planar)	8–9	32 H_2O (28+4)[2] 17–19 Ar, N_2, O_2 7·5 I_2 6·5 SF_6 5·8 C_2F_6 5·6 cyclopentane 5·4 benzene 4·6 toluene 4·5 n-C_5H_{12} 4·1 cyclohexane 4·1 perfluoro-cyclobutene 4·1 $C_2F_4Cl_2$ 3·5 n-C_7H_{16} 3·4 C_3F_8 2·9 n-C_4F_{10} 2·8 iso-octane 2·3 perfluoro-methylcyclohexane 2·1 perfluoro-di-methylcyclohexane

The mobility of guest species saturating the channel systems of zeolites may be considerably less than it is in their bulk liquids. In hydrated chabazite, gmelinite, and heulandite, energy barriers for self-diffusion of zeolitic water are considerably greater than they are in liquid water, and self-diffusion coefficients extrapolated to −2°C are intermediate between that in ice and that in supercooled liquid water [44]. On such evidence, zeolitic water is intermediate in mobility between bulk liquid and ice. Measurements of the integral thermal entropy of zeolitic ammonia in various ion-exchanged forms of a Sieve X—a different approach to the problem—lead to the conclusion that zeolitic ammonia is appreciably more restricted in its vibrations, librations, and Brownian movement in the zeolite than in the bulk liquid [45]. These facts are not surprising in view of steric restrictions imposed by the narrow channels in the zeolite, and of the strong and rather inhomogeneous binding forces acting on both H_2O and NH_3.

Factors governing molecular sieving

The versatility and range of application of molecular sieve sorbents has recently been described in another article [46] and will not be considered here. However, it is of interest to summarize the factors, often controllable, that govern the sifting of molecular mixtures using zeolite crystals as filters. The term 'molecular sieving' is quite appropriate in this context. Thus, figure 9 [47] represents in a formal way the surface of a mordenite crystal cut normal to the *c*-axis. At each corner of the line diagram is an Al or Si atom; oxygens centred near but not on the mid-points of all edges are not shown. The complete uniformity of the openings to the channels is evident and is fully analogous, on a molecular scale, to the mesh of a mechanical sieve. Molecules too large or of the wrong shape to pass the mesh are turned back at the surface, but smaller molecules pass the mesh readily and can be copiously sorbed.

Apart from the special ways of representing zeolite structures already described, one may recognize in them 4-, 5- 6-, 8-, 10-, and 12-rings and one may construct any zeolite whatever by appropriate linkings of one or more kinds of these rings. Even small molecules which are to pass into the body of the crystal will never be small enough to push their way through 4- and 5-rings, but will pass through 6-, 8-, 10-, and 12-rings with increasing ease. The largest mesh in figure 9 is composed of 12-rings, and we have considered also certain 6- and 8- and 12-ring filters when discussing table II. The 12-ring in figure 9, shown there only in projection, is puckered, and in fact the

---

[1] These dimensions refer to the 8-ring on a dehydrated Ca-rich specimen [37].

[2] The division of water molecules in this way represents 4 as full occupation of the smaller cubo-octahedral cages also present in these two structures, as in basic sodalite. Full occupation may not always happen because $NaAlO_2$ may sometimes occupy cubo-octahedral cavities at least in Sieve A [24].

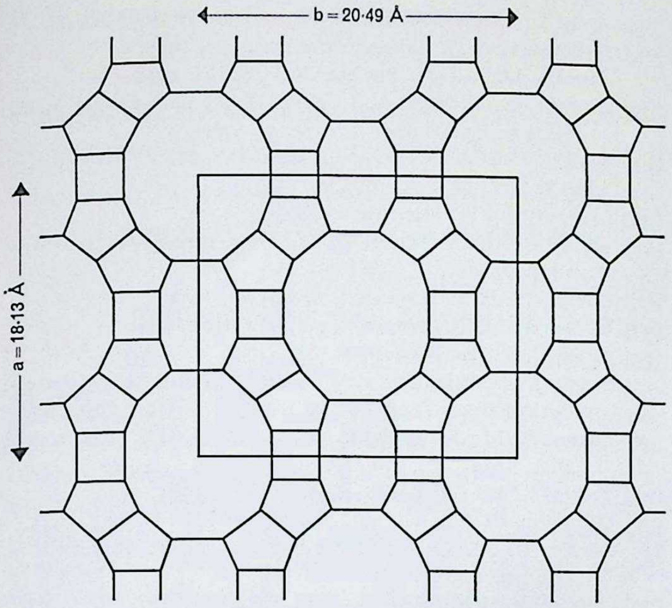


Figure 9 A section of the mordenite structure in the *ab* plane, showing the sieve-like character of the surface. Al or Si atoms are located at each apex; oxygens are sited near, though not on, the midpoints of all edges [47].

free dimensions (figure 5) are 5·9 (minimum) and 7·1 Å (maximum), and are thus appreciably smaller than the 12-ring mesh that gives access into faujasite.

Thus, all 12-ring windows are not equivalent, owing to varying degrees of non-planarity or puckering. Exactly the same is true of other windows composed of smaller numbers of atoms. Figure 10 [33] illustrates the non-planarity and the free dimensions of the 8-ring windows in Linde Sieve A, chabazite, erionite, Na-P1 and levynite. Only Sieve A has planar windows, and accordingly one finds that this zeolite freely admits *n*-paraffins. Some of the other zeolites will also do this, but the ring distortion restricts the rate of diffusion considerably. Thus structures all having 8-ring windows will in practice have a range of molecular sieve behaviours on account of ring distortion alone.

Another important factor is the lack of total rigidity of the frameworks. In the only structure on which the effect of dehydration has been examined (chabazite) it is found that the minimum and maximum free dimensions of the 8-ring window change from those in figure 10 (3·7 and 4·1 Å) to those in table II (3·1 and 4·4 Å). Such changes will differ for the different zeolites and must help to determine the gauge of the mesh and introduce variety into this gauge. Moreover, if the windows do change dimensionally in this way on dehydration, they may also change dimensionally when a guest molecule is pushing its way through them. Further, neither the oxygen atoms that form the inner rim of the windows nor the guest molecules are incompressible. Thus, we would in general expect that guest molecules of somewhat larger van der Waals dimensions than the windows could pass through them, albeit with rapidly increasing difficulty as the discrepancy between their dimension and that of the window increased. This difficulty would be expressed as an increased energy barrier, E, and a more negative entropy of activation, in the Arrhenius equation for the diffusion coefficient, $D=D_0\exp(-E/RT)$. The entropy term will appear in the coefficient D_0.

Because of the sensitiveness of E to the molecular dimensions of the guest molecule, and because the temperature T also appears in the exponential term, the sensitiveness of molecular sieving can sometimes be greatly increased by lowering the temperature. For example, at −78°C, both oxygen and nitrogen are readily sorbed by the sodium form of Sieve A. At −183°C, however, only the slightly smaller oxygen molecule is sorbed. Similarly the small differences in size of argon and oxygen are sufficient to enable these two molecules to be separated from their mixtures using Ca-mordenite at −183°C. Only oxygen is then sorbed [48].

The cations are present in the same channels as the guest molecules. They are often recessed into 6-ring windows, which do not normally function as the windows giving access to the channel system. They may also be located in subsidiary polyhedra not themselves able to hold guest molecules, such as the hexagonal prisms in faujasite or Sieve X. Such cations do not hinder the migration of guest molecules. Other cations can still remain, however, and if these are in or near the windows they then become 'sentinels' that control access to the channel systems in a decisive way. One may moderate this influence in three ways [11]:

1. By changing the size of the cations through exchange ($K^+ \rightleftharpoons Na^+$, for example).
2. By changing the number of the cations through exchange ($2Na^+ \rightleftharpoons Ca^{++}$, for example).
3. By changing the number of the cations through synthesis ($NaAl \rightleftharpoons Si$, for example).

Thus in the case of chabazites of oxide compositions $M_2O.Al_2O_3.4SiO_2$, or thereabouts, the K-form sorbs only small polar molecules such as H_2O and NH_3; the Na-form sorbs also permanent gases; while the Ca-form sorbs even *n*-paraffins. The K-, Na-, and Ca-forms of Sieve A behave similarly. If, by direct synthesis, one prepares more and more aluminous near-chabazites, even the Na-form loses its ability to sorb permanent gases, presumably because the structure

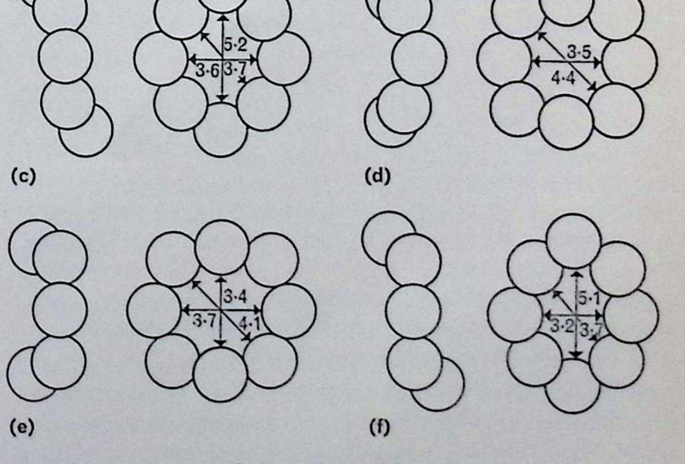


Figure 10 Diversity of shape and of free dimensions in some 8-rings in various hydrated zeolites [33]. (*a*) zeolite A, (*b*) chabazite, (*c*) erionite, (*d*) Na-P1, (*e*) gmelinite, (*f*) levynite. Rings of linked tetrahedra always have their inner edges formed of oxygen atoms in contact, and equal in number to the tetrahedra. It is these oxygen atoms which are shown in the diagram. (All dimensions in Å.)

becomes further enriched in Na^+ ions, more and more of which then act as sentinels and prevent access of guest molecules.

These effects are not always in one direction. Among the mordenites the Na-form admits guest molecules of larger dimensions than the Ca- or Ba-forms, although there are fewer alkaline-earth ions than Na^+ ions after exchange ($2Na^+ \rightleftharpoons Ca^{++}$). Also, the Ca-form of Sieve X, although still a very open structure, is less open than the Na-form, which contains twice as many cations.

A further means of moderating the molecular sieve behaviour of a given zeolite has proved possible. One may add controlled amounts of strongly adsorbed modifier molecules which are relatively immobile at the temperature of the subsequent sorptions, and which can therefore act as additional barriers in the channel systems. Water, ammonia, and methylamine, for example, have proved effective modifiers of sorption of permanent and rare gases by mordenite [11]. The modifying action can be selective as between molecules of different dimensions.

References

[1] Maggs, F. A. P. *Nature, Lond.*, 169, 793, 1952.

[2] Metcalfe, J. E., Kawahata, M., and Walker, P. L. *Fuel*, 42, 233, 1963.

[3] Dollimore, D. and Heal, G. R. *Trans. Faraday Soc.*, 59, 2386, 1963.

[4] McBain, J. W. 'The Sorption of Gases by Solids'. Chapter v. Routledge and Kegan Paul, London. 1932.

[5] Barrer, R. M. *Proc. roy. Soc., A*, 167, 392, 1938.

[6] *Idem. J. Soc. chem. Ind., Lond.*, 64, 130, 133, 1945.
Barrer, R. M. and Belchetz, L. *Ibid.*, 64, 131, 1945.

[7] *Trans. Brit. Ceram. Soc.*, 56, 155, 1957.

[8] Barrer, R. M. *Nature, Lond.*, 159, 508, 1947.

[9] Kerr, G. T. and Kokotailo, G. T. *J. Amer. chem. Soc.*, 83, 4675, 1961.
Kerr, G. T. *Science*, 140, 1412, 1963.

[10] Barrer, R. M. and Denny, P. J. *J. chem. Soc.*, 971, 1961.
Barrer, R. M., chapter 6, table 24, in 'Non-stoichiometric Compounds' (edited by L. Mandelcorn). Academic Press, New York. 1964.

[11] *Idem. Trans. Faraday Soc.*, 45, 358, 1949.
Barrer, R. M. and Rees, L. V. C. *Ibid.*, 50, 852, 1954.
Barrer, R. M. and Baynham, J. W. *J. chem. Soc.*, 2892, 1956.

[12] Reed, T. B. and Breck, D. W. *J. Amer. chem. Soc.*, 78, 5972, 1956.
Broussard, L. and Shoemaker, D. P. *Ibid.*, 82, 1041, 1960.

[13] Bergerhoff, G., Baur, W. H., and Nowacki, W. *Neues Jb. Miner., Mh.*, 9, 193, 1958.

[14] Meier, W. M. *Z. Kristallogr.*, 115, 439, 1961.
Gottardi, G. and Meier, W. M. Miner. Soc. Amer. (3rd Gen. Mtg, Int. Miner. Assoc.), Special Paper 1, 1963.
Kerr, I. S. *Nature, Lond.*, 197, 1194, 1963.

[15] Dent, L. S. and Smith, J. V. *Ibid.*, 181, 1794, 1958.

[16] Staples, L. W. and Gard, J. A. *Miner. Mag.*, 32, 261, 1959.

[17] Sadanaga, R., Marumo, F., and Takeuchi, Y. *Acta Crystallogr.*, 14, 1153, 1961.
Steinfink, H. *Ibid.*, 15, 644, 1962.
Smith, J. V. and Rinaldi, F. *Miner. Mag.*, 258, 202, 1962.

[18] Barrer, R. M., *Proc. chem. Soc.*, 99, 1958.

[19] Ames, L. L. *Amer. Miner.*, 46, 1120, 1961; 47, 1067, 1962.

[20] Deffeyes, K. S. *J. sediment. Petrol.*, 29, 602, 1959.
Coombs, D. S. *Trans. roy. Soc. N.Z.*, 82, 65, 1954.

[21] Coombs, D. S., Ellis, A. J., Fyfe, W. S., and Taylor, A. M. *Geochim. et cosmochim. Acta*, 17, 53, 1959.

[22] Walker, G. P. L. *Miner. Mag.*, 32, 503, 1960.
Idem. J. Geol., 68, 515, 1960.

[23] Loewenstein, W. *Amer. Miner.*, 39, 92, 1954.

[24] Barrer, R. M. and Meier, W. M. *Trans. Faraday Soc.*, 54, 1074, 1958.

[25] Taylor, W. H. *Proc. roy. Soc., A*, 145, 80, 1934.

[26] Meier, W. M. *Z. Kristallogr.*, 113, 430, 1960.

[27] Smith, J. V. and Rinaldi, F. (See [17].)
Smith, J. V. Miner. Soc. Amer. (3rd Mtg., Int. Miner. Assoc.), Special Paper 1, 281, 1963.

[28] Barrer, R. M., Bultitude, F. W., and Kerr, I. S. *J. chem. Soc.*, 1521, 1959.

[29] Barrer, R. M. and White, E. A. D. *Ibid.*, 1561, 1952.

[30] Nithollon, P., *Publ. sci. tech. Ministère Air*, No. N.T.53, 1955.

[31] Barrer, R. M. Colston Papers, p. 6, vol. x. Butterworths, London. 1958.

[32] *Idem. Trans. Faraday Soc.*, 40, 555, 1944.

[33] Barrer, R. M. and Kerr, I. S. *Ibid.*, 55, 1915, 1959.

[34] Halstead, P. E. and Moore, A. E. *J. appl. Chem.*, 12, 413, 1962.

[35] Barrer, R. M. and Wasilewski, S. *Trans. Faraday Soc.*, 57, 1140, 1961.

[36] Meier, W. M. and Kokotailo, G. T. 6th Int. Union of Crystallography, 1963; *Z. Kristallogr.* (In press.)

[37] Smith, J. V. *Acta Cryst.*, 15, 835, 1962.

[38] Barrer, R. M., Bultitude, F. W., and Sutherland, J. W. *Trans. Faraday Soc.*, 53, 1111, 1957.
Barrer, R. M. and Rees, L. V. C. *Ibid.*, 55, 992, 1959.

[39] Barrer, R. M. and Reucroft, P. J. *Proc. roy. Soc., A*, 258, 431, 1960.
Barrer, R. M. and Sutherland, J. W. *Ibid., A*, 237, 439, 1956.

[40] Barrer, R. M. and Gibbons, R. M. *Trans. Faraday Soc.*, 59, 2569, 1963.
Barrer, R. M. and Reucroft, P. J. *Proc. roy. Soc., A*, 258, 449, 1960.
Barrer, R. M. and Stuart, W. I. *Ibid., A*, 249, 464, 1959.

[41] Kington, G. L. and Macleod, A. C. *Trans. Faraday Soc.*, 55, 1799, 1959.

[42] Kerr, G. T. and Johnson, G. C. *J. phys. Chem.*, 64, 381, 1960.
Barrer, R M. and Brook, D. W. *Trans. Faraday Soc.*, 49, 940, 1953.
Addison, W. E. and Barrer, R. M. *J. chem. Soc.*, 757, 1955.
Frilette, V. J., Weisz, P. B., and Golden, R. L. *J. Catalysis*, 1, 301, 1962.
Weisz, P. B., Frilette, V. J., Maatman, R. W., and Mower E. B. *Ibid.*, 1, 307, 1962.
Bassett, D. W. and Habgood, H. W. *J. phys. Chem.*, 64, 769, 1960.

[43] Stamires, D. N *J. chem. Phys.*, 16, 3174, 1962.

[44] Barrer, R. M. and Fender, B. E. F. *J. Phys. Chem. Solids*, 21, 12, 1961.

[45] Barrer, R. M. and Gibbons, R. M. *Trans. Faraday Soc.*, 59, 2875, 1963.

[46] Barrer, R. M. *Brit. chem. Engng.*, 7, 267, 1959.

[47] Eberly, P. E. *J. phys. Chem.*, 67, 2404, 1963.

[48] Barrer, R. M. and Robins, A. B. *Trans. Faraday Soc.*, 49, 807, 1953.

Some applications of the Mössbauer effect

P. B. Moon and D. A. O'Connor

The spectral lines of gamma rays, as emitted from a nucleus, normally have some width, but R. L. Mössbauer reported in 1958 that this width could be extremely small under certain conditions. The almost monoenergetic gamma rays produced by the Mössbauer effect have been used in determining the mass of photons, in reducing the limits of any possible 'aether drift' to very low levels, and in measuring the intrinsic magnetic field of crystals. This article describes this effect and its applications, both present and future, that include the simplification of crystal-structure analysis.

This article is concerned with gamma rays emitted when an atomic nucleus changes from a state of relatively low internal energy to the lowest possible state. The energies of the individual gamma rays would ideally be just equal to the difference of energy between the upper, excited, state and the lowest, ground, state, but the indeterminacy principle tells us that there is an uncertainty $h/2\pi\tau$ in the energy of the level and of the gamma rays, where h is Planck's radiation constant and τ the average time spent in the excited state. Thus individual gamma rays do not all have exactly the same energy; in optical language, the gamma-ray line must have some width.

Nuclear physicists like to express both the energy and the spread of energy in electron-volts (eV), and in these terms the intrinsic width (ΔE) of the line is about $(6\times10^{-16})/\tau$ eV, where τ is the average lifetime, in seconds, of the particular excited state in question. Since the actual energies, E, of the gamma rays lie between about 10^4 and 10^5 eV, and the lifetimes range from about 10^{-7} to 10^{-9} second, the fractional widths $\Delta E/E$ are extremely small; for the iron-57 gamma ray of energy $1{\cdot}4\times10^4$ eV and lifetime 1×10^{-7} sec, this fraction is about 4×10^{-13}. This is to be compared with fractional widths of the order 10^{-6} for optical lines and 10^{-4} for X-ray lines, and it represents a degree of homogeneity not approached by any other form of radiation.

Normally, this narrowness cannot be achieved. The source nuclei are moving in random directions with the thermal velocities of their atoms, usually 10^4 cm per sec or more. The Doppler effect of these motions spreads the line by a fractional width measured by the ratio of the source speed to the speed of light (3×10^{10} cm per sec). It is thus of the order of 10^{-6}. Moreover, the nucleus recoils when it emits the gamma ray: this takes energy and shifts the gamma-ray line by an amount which is usually comparable with the Doppler spread.

Six years ago, R. L. Mössbauer reported [1] that if the source nuclei are embedded in a cold crystalline solid, a fraction of the gamma rays are emitted without Doppler spread and without recoil shift. The recoil momentum given to the crystal is the same as that given to a free atom, but if the atoms in the crystal are rigidly bound, then the crystal recoils as a whole and the recoil energy given to the whole crystal is negligible. However, the binding of atoms in crystals is elastic rather than rigid and there exists the possibility of the transference of some of the nuclear energy to vibrational energy of the atoms in the crystal lattice. The vibrational modes of the crystal are quantized and the probability of their excitation depends on the amount of recoil energy being offered, on the energy of the vibrational states, and on the crystal temperature. It turns out that there is a finite probability that the crystal will recoil as a whole without exchange of energy with the gamma ray.

Mössbauer's discovery of recoilless gamma-emission was made with the simultaneous discovery of recoilless absorption, in a foil containing nuclei of the same type (of course, in their ground state). The recoilless gamma rays have exactly the right energy to raise these nuclei to the upper energy level on those occasions when the crystal in which they are embedded declines to exchange any energy with the nucleus.

Diagrammatically, the apparatus is as in figure 1(b); the nuclear processes are shown in figure 1(a). If the two crystalline environments are identical, the emission and absorption lines are exactly matched in energy; but in any event one line may be 'scanned' by the other, using the Doppler effect of a relative motion of source and absorber. Figure 1(c) shows how the counting rate will change as this scanning occurs. It should be noted that, because only a fraction of the gamma rays are emitted, or absorbed, without recoil the dip is not usually very deep. The lower the gamma-ray energy and the more rigid the binding of the atoms in the crystal, the deeper is the dip.

Most gamma rays are inherently incapable of showing the Mössbauer effect, either because they do not go to the nuclear ground state, or because the ground state is itself unstable, so that no resonant absorber can be had. Even when it is possible in principle, it seldom happens that the gamma-ray energy is low enough to give an appreciable recoilless fraction, even at the lowest temperatures. Of the thousands of known gamma rays, just three can give relatively large effects at room temperature: the iron line already mentioned, and lines in isotopes of tin and thulium, ^{119}Sn and ^{169}Tm, respectively. Mössbauer's discovery was made with a selective absorption in ^{191}Ir amounting to less than 1 per cent and requiring temperatures far below 0°C.

Equipment

The apparatus required to investigate the recoilless absorption effect consists of a source, an absorber, a gamma-ray detector, and a device to produce relative motion between the source and absorber (figure 1). The detector can be either a scintillation counter or a gas proportional counter. In either case, the counter must be adjusted electronically to register only those pulses that correspond to gamma rays of the required nuclear transition energy. Gamma rays emitted with or without recoil loss of energy will be detected with the same efficiency. The motion-producing device is normally required to produce relative velocities giving a Doppler shift of the order of the nuclear line width. The line width, expressed in terms of shift velocity and

P. B. Moon, *M.A., Ph.D., F.R.S.*
Was borni n 1907 in London and educated at Leyton County High School and Sidney Sussex College, Cambridge. He has lectured at Imperial College and at the University of Birmingham, and was appointed to the Poynting Chair of Physics at the University of Birmingham in 1950.

D. A. O'Connor, *B.Sc.*
Was born in London in 1927 and educated at Leyton County High School; he took an external degree at London University in 1947. His first research was at the National Institute for Medical Research, London, and he was later at the Institute for Nuclear Research, Warsaw. Since 1960, he has been a lecturer in the Physics Department of the University of Birmingham.

Table I

Isotope	Gamma-ray energy	Line width		Fraction of radiation emitted or absorbed without recoil in elemental form at temperatures of:	
	keV	eV	mm per sec	300°K	4°K
^{57}Fe	14·4	$4{\cdot}5\times10^{-9}$	0·1	0·7	0·9
^{119}Sn	24·0	$2{\cdot}4\times10^{-8}$	0·3	0·05	0·71
^{161}Dy	25·7	$1{\cdot}6\times10^{-8}$	0·2	0·13	0·80
^{169}Tm	8·4	$1{\cdot}1\times10^{-7}$	4·0	0·7	0·95
^{197}Au	77·0	$2{\cdot}4\times10^{-7}$	1·0	$<10^{-4}$	0·20

energy, and the gamma-ray energy and the recoilless fractions of the most popular Mössbauer nuclides are given in Table 1.

As we shall see later, there are several kinds of interaction of the nucleus with the crystal lattice which, because the energy associated with them is larger than a nuclear line width, lead to modifications of the absorption and emission line patterns. The single line considered above can in a different crystalline environment be broadened, shifted from zero velocity, or split into several discrete lines. ^{57}Fe in metallic iron gives a pattern of lines in which the extreme lines are shifted in energy from the centre of the pattern by as much as 100 times the single line width. To examine these lines, it is necessary to move the source or absorber at velocities 100 times the velocity corresponding to the line width, and with a constancy much better than a line width. For this sort of work with ^{57}Fe and ^{119}Sn, motion-producing devices have to be constructed that will cover a large range of velocities, from 0·01 mm per sec to 10 mm per sec, and reproduce these velocities with a precision of better than 0·01 mm per sec. The apparently trivial problem of moving a small object at a known velocity turns out to be difficult technically, as the instantaneous velocity must be known accurately. Mechanical devices based on rotating screw-threads and cams can be made to give accurately known average velocities, but since they are prone to vibration, they do not always give a reproducible instantaneous velocity. Other devices which are used are based on transducers of the moving-coil loudspeaker type. These can be made relatively free from extraneous vibrations but present problems of velocity measurement. In practice, both types of device are in use, to produce both motion of constant velocity and motions in which the velocity is swept through a known range. The latter type of device is used with a synchronized multi-channel counting device to produce a complete 'velocity spectrum'.

Sources and absorbers

Since the nuclear resonance is so sensitive to crystalline environment, it is necessary to ensure that the source and the absorber each exist in a uniform environment. The production of a uniform source can be particularly difficult, since the precursor of the source nucleus can be different chemically from the source itself. For example, the 14·4 keV excited state of ^{57}Fe is formed as a result of the decay of the 270-day ^{57}Co isotope to the 136 keV state of ^{57}Fe, which in turn decays by gamma emission to the 14·4 keV state with a half-life of less than 5×10^{-8} sec. A strong source of the ^{57}Fe 14·4 keV radiation will therefore contain appreciable amounts of residual cobalt. If the host lattice is of iron or a compound of iron, then the cobalt is present as an impurity and the cobalt atoms may not necessarily be at normal lattice sites or even in identical environments. The rearrangement of electronic shells caused by the change in atomic number apparently takes place so rapidly that it does not in itself affect the properties of the 14·4 keV state. The precursor of the 24 keV state of ^{119}Sn is a long-lived 89 keV state of ^{119}Sn itself, so this source is intrinsically uniform.

The resonant absorber must contain an appreciable proportion of the resonant nuclei in the ground state. For both ^{57}Fe and ^{119}Sn, the normal isotopic abundance of the relevant isotopes is usually sufficient to produce an observable Mössbauer effect. However, it is often necessary to enrich the sample in the resonant isotope.

In most studies of the Mössbauer effect in crystals, the substance studied is used as an absorber, and a source is chosen that gives a single resonance line with a large recoilless fraction. The oxide of tin, SnO_2, is a suitable ^{119}Sn source. Single line sources of the ^{57}Fe radiation are prepared by incorporating the ^{57}Co precursor in non-magnetic stainless steel, or in other metals such as copper, titanium, and platinum. Stainless steel normally contains inactive cobalt, but in the other cases the ^{57}Co is an impurity; if its concentration is very low, then all the cobalt nuclei will be in identical environments.

Applications to tests of relativity

Measurements of the transmission of a resonant absorber can be used to determine a small difference in energy between the source

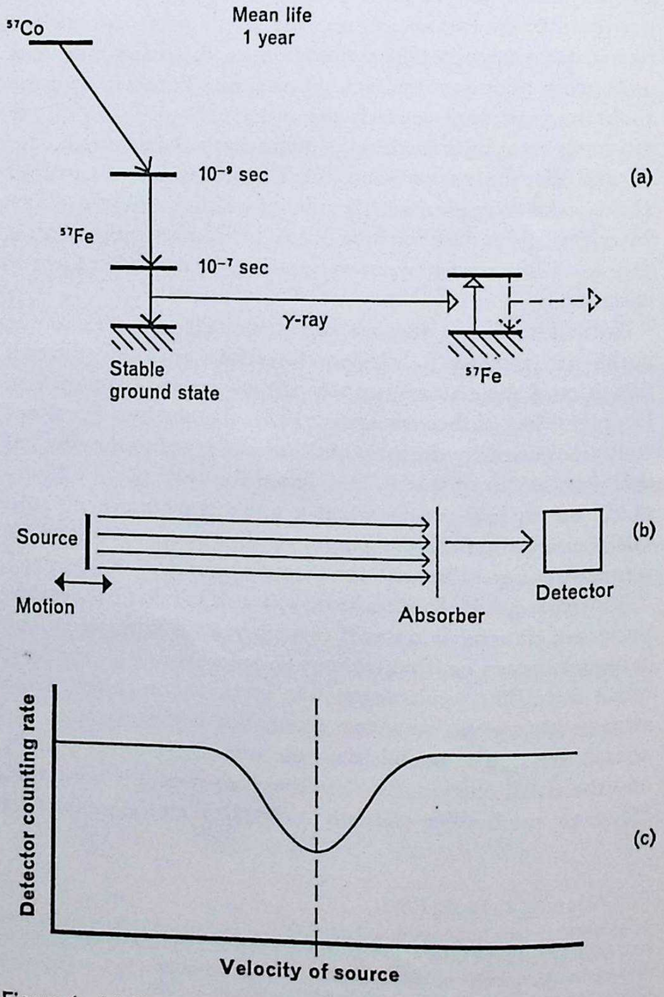


Figure 1 (*a*) Decay scheme for the ^{57}Fe system. The absorber transition shown in dotted lines represents the re-emission of a gamma ray which follows a fraction of the absorption processes. (*b*) Arrangement of apparatus corresponding to (*a*). (*c*) Typical curve showing dependence of detector counting rate on source velocity.

emission line and the absorber absorption line. It is clear from the shape of the transmission curve of figure 1 that the greatest sensitivity to a shift in the source or absorber lines will be obtained when the emission line is already some way out in the wing of the absorption curve. This fact has been exploited in the most delicate application of the Mössbauer effect so far made, the verification of the effective mass of photons through the change of their energy as they move from one height to another in the earth's gravitational field.

R. V. Pound and G. A. Rebka [2] oscillated an ^{57}Fe source with a sinusoidal motion along a vertical line, the source being 74 ft from the absorber and detector, either at the top of a tower, or at the bottom with the absorber and detector at the top.

Gamma-ray counts occurring during the quarter-cycles of greatest upward and downward velocity, that is, when the emission line was shifted to the steep slopes to either side of the centre of the absorption line, were separately recorded and added over long periods of time. Their differences corresponded to a relative displacement of line centres by 19·7 ± 0·8 parts in 10^{15} with the source at bottom, and 15·5 ± 0·8 with the source at top: the mean of 17·6 represents a difference between the characteristics, chemical or physical, or both, of source and absorber. The difference of 4·2 ± 1·1 represents twice the gravitational shift for a 74 ft vertical path. The first set of such experiments agreed with theory within the experimental accuracy of about 10 per cent; later experiments, understood to be still in progress, have much improved this accuracy.

That a photon gains energy from a gravitation field as if it were a particle of mass $h\nu/c^2$ is, of course, a prediction of relativity theory. Historically, the theory was required to account for the non-observance of any effects of motion of the earth through the 'aether' that in nineteenth-century physics was supposed to be the medium through which light travels. Although the existence of an aether is irrelevant to the theory as now understood, it is clearly of interest to check the absence of aether drift to higher accuracy as more precise tests become available. The original optical experiments of Michelson and Morley set a limit of several kilometres per second; a recent measurement by T. S. Jaseja, A. Javan, J. Murray, and C. H. Townes [3], using a laser, reduced it to about 10 metres per sec; a Mössbauer effect measured by D. C. Champeney, G. R. Isaak, and A. M. Khan [4] has shown the east–west component of such a drift at Birmingham to be less than about 3 metres per sec.

In this experiment, source and absorber were mounted diametrically opposite on a high-speed rotor: an east–west drift would, curiously enough, produce a north–south asymmetry. To make the test of this as sensitive as possible, the source atoms were in a matrix different from the absorber, the choice being such that the shift was about half a line width.

Applications to crystal physics and chemistry

The relativity applications above concern small shifts of energy of gamma-ray photons in flight; we shall now mention some applications that involve small changes of the energy of either emission or absorption lines, arising from changes in the environment of, respectively, the emitting or absorbing nuclei. The intrinsic magnetic field of a ferromagnetic or paramagnetic crystal will split the lines into several components: the number of components depends on the spins of the nucleus in its upper and ground states: the spacing is proportional to the field at the nucleus and also depends on the spins and magnetic moments of these states.

For both ^{57}Fe and ^{119}Sn, the upper state splits into four states of slightly different energy, and the ground-state into two. However, two of the eight possible transitions are forbidden by the 'selection rules' that are well known to atomic and nuclear spectroscopists, so the emission and absorption lines are each split into six components. Figure 2 shows the form of the motion spectrum obtained from an unsplit ^{57}Fe source scanned over a magnetically split ^{57}Fe absorber (or, of course, from a split source and an unsplit absorber). If both are split, the situation is more complicated, but is still capable of analysis. This analysis gives values for the intrinsic magnetic field of the nucleus.

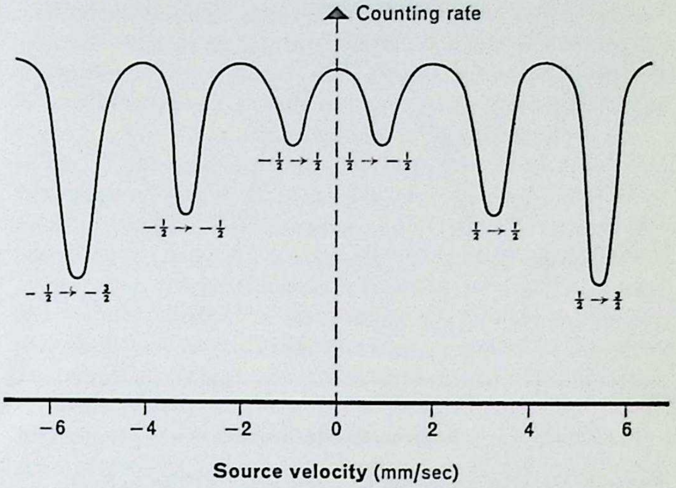


Figure 2 Velocity spectrum obtained when an unsplit ^{57}Fe source is used to scan the resonances of a magnetically split ^{57}Fe absorber. The absorber resonance lines correspond to the transitions between the ground and excited spin states indicated.

The first outstanding measurement of this class was that of the magnetic field at the nuclei in metallic iron, which was shown by S. S. Hanna, J. Heberle, C. Littlejohn, G. J. Perlow, R. S. Preston, and D. H. Vincent [5] to be about 300 000 oersteds. This is more than ten times the average field, and what is remarkable is that its magnitude and even its sign—parallel or antiparallel to the average field in the locality—were previously quite unknown. Many measurements have since been made of the magnetic fields at iron, or tin, nuclei in various alloys and compounds.

Except when, as in a cubic crystal, the nucleus is a centre of symmetry of the neighbouring atoms, there will be a gradient of electric field over the region of the nucleus. If then, as for both ^{57}Fe and ^{119}Sn, the ground state and the excited state have electric quadrupole moments (in more pictorial language, if they are ellipsoidal as regards their distribution of electric charge), different directions of the nuclear axis give different energies; so the energy-levels, and thus the emission and absorption lines, may be split.[1] A measurement of this splitting enables the field gradient to be determined if the quadrupole moment is known, or vice versa. Figure 3 shows the electric-quadrupole doublet of ^{119}Sn in $(CH_3)_3SnCl$ scanned with the unsplit line of SnO_2 [6].

A more subtle effect is known as the isomeric shift or the chemical shift and depends upon the different volumes of the nucleus in its two states and the different electron densities in the neighbourhood of the nucleus when the atom in question is built into two different crystals. The *s* electrons (those that have no orbital

---

[1] Because the electric quadrupole moment has only a direction and not a polarity, the electrostatic splitting gives only half the number of different energy levels, in both the ground and excited states, that the magnetic splitting gives. The number of components in the spectrum line is much reduced: for ^{57}Fe and ^{119}Sn the sextuplet becomes a doublet.

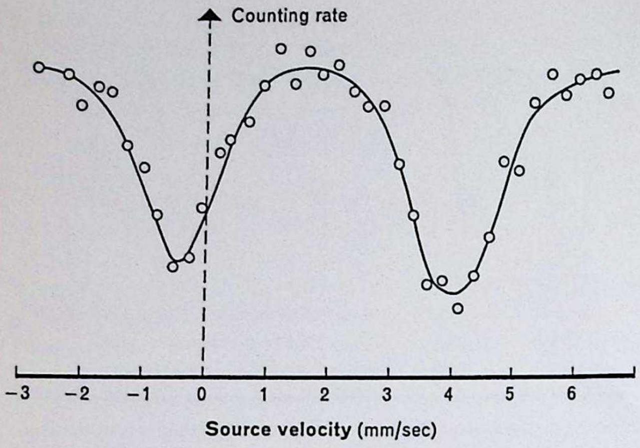


Figure 3 Velocity spectrum obtained when a ^{119}Sn in SnO_2 source is used to scan the quadrupole split ^{119}Sn in $(CH_3)_3SnCl$.

angular momentum) have a non-zero probability of being within the nuclear volume. The probability of their being within this volume reduces the potential energy by an amount proportional to the nuclear charge, the nuclear radius, and the time-average of the electron concentration. If the nuclear radius R is different for the ground and excited states, the gamma-ray energy will be slightly changed, by an amount proportional to the change in R and to the electron concentration. This shows as a small shift in the position of a Mössbauer line (or of the components of a split line) as between one chemical environment and another. Most of such comparisons of *s*-electron density, which in conjunction with the electric splitting described above give information about the character of the bonding of the atom with its neighbours in the crystal, have been made between various compounds of tin (table II). Valencies of two and four, and various proportions of covalent bonding, give a wide variety of situations, some of which have been studied by A. J. F. Boyle, D. St P. Bunbury, and C. Edwards in Manchester [7], N. N. Delyagin and V. J. Goldanskii in the Soviet Union [8], and M. Cordey-Hayes in Birmingham [9].

Table II

Bond	Electronegativity difference	Shift of ^{119}Sn line relative to Sn-Sn (mm per sec)
Sn-F	2·0	2·6
Sn-O	1·7	2·2
Sn-Cl	1·2	1·3
Sn-Br	1·0	1·0
Sn-S	0·7	0·8
Sn-I	0·6	0·3
Sn-Sn	0·0	0·0

The electronegativity difference quoted is based on chemical data. Electronegative ions attract *s*-electrons away from the tin atoms, and this results in a shift in the Mössbauer resonance line.

Another type of investigation in a borderland between physics and chemistry has been initiated by S. L. Ruby, P. A. Flinn, and W. L. Kehl at Westinghouse [10]. They adsorbed ^{57}Fe atoms on surfaces of Al_2O_3 powder and have begun to investigate the surface fields through their effect on the ^{57}Fe absorption spectrum.

Applications to crystal and molecular structure analysis

Since recoilless radiation can be obtained only from low-energy gamma sources, the wavelengths are comparable with crystal cell dimensions. For example, ^{57}Fe emits radiation of 0·86 Å wavelength. It follows that this radiation can be used in crystal-diffraction experiments in place of the usual X-rays. A radioactive source of moderate strength and cost emits far fewer photons than does an X-ray tube, but the recoilless gamma rays are superior to X-rays in one important respect: the definition of their energy and wavelength is very much better. Whereas the smallest energy spread attainable for X-rays is about 0·1 eV, the energy width of the 14·4 keV narrow line from the excited state of ^{57}Fe is $4{\cdot}5 \times 10^{-9}$ eV. Even with thermal Doppler broadening, a typical low-energy gamma line is only about 0·02 eV broad.

When photons are scattered by a crystal, some are scattered without change in energy and give rise to the Bragg diffraction peaks characteristic of the crystal structure. In addition, some may be scattered with change in energy, giving rise to a diffuse background to the peaks which itself has maxima at the peaks. In the neighbourhood of the diffraction peaks, the diffuse background corresponds to the interaction of the photons with very long wavelength thermal waves with energy as low as 10^{-8} eV. This energy change is beyond the reach of X-ray spectrometry but can be detected with recoilless gamma rays. If some of the gamma rays suffer an energy change on being

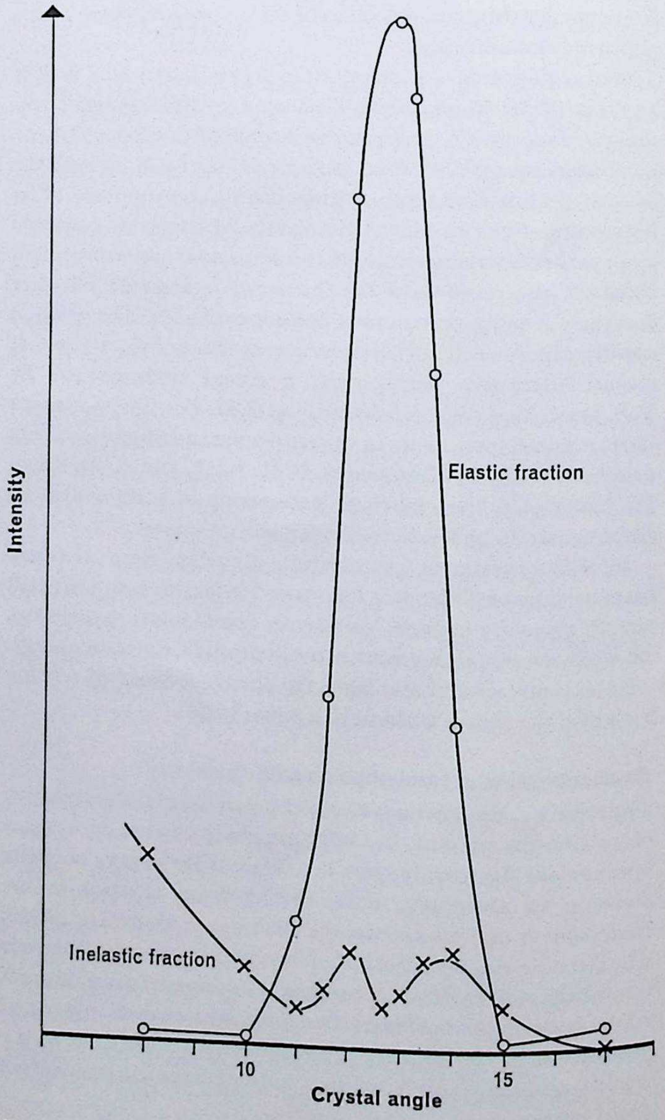


Figure 4 The elastic and inelastic components of 14·4 keV gamma radiation Bragg-reflected by the (200) planes in a lithium fluoride single crystal.

scattered by a crystal, then the scattered beam will suffer a smaller absorption in a resonantly absorbing foil than would the incident beam. This is the basis of the method used by D. A. O'Connor and N. M. Butt [11] to demonstrate experimentally for the first time the peak in the diffuse scattering by a single crystal of lithium fluoride at a Bragg diffraction peak (figure 4).

The significance of this experiment lies in the fact that it is the measurement of the strictly elastic part of the diffraction peaks which is the basis of crystal structure determination. The X-ray crystallographer in his measurements of the intensities of these peaks inevitably includes some part of the inelastic diffuse scattering, and this will often amount to a few per cent of the total. Theoretical corrections for this effect themselves depend on a knowledge of details of the crystal structure, and hitherto no experimental method of correction has been available.

Another feature of Mössbauer gamma rays may in the future have applications to structural analysis. We refer to the fact, demonstrated experimentally by P. J. Black and P. B. Moon [12], that the resonant elastic scattering of gamma rays by nuclei can interfere, in the optical sense of the word, with their elastic scattering by atoms, and that the phase, together with the amplitude, of the nuclear resonant scattering can readily be varied.

The determination of molecular structures by X-ray diffraction can be described as a guessing game in which, with some initial chemical and other physical evidence to help him, the player has a set of numerical clues, the intensities of the Bragg peaks. Reasonable guesses of the structure must be made, the corresponding intensities calculated, and their fit with the experimental values tested. For complicated structures, the process is greatly assisted by, and indeed may be impracticable without, some information about the phase of the wave scattered from one part of the molecule as compared with the phase of the scattering from the rest. The device adopted is to build a heavy atom (one that contributes appreciably to the total scattering) into the molecular structure, and to compare the intensities of reflections for X-rays that have energies below and above a critical absorption energy of this atom. The phase of part of its scattering is 180 degrees different on the two sides.

The Mössbauer effect offers, in principle, an alternative method of phase-investigation. If, for example, the molecule contains iron, the scattering of the ^{57}Fe gamma ray by the nucleus will change phase by 180 degrees in response to the very small change in energy required to go from one side of the resonance to the other. Since the nuclear scattering amplitude is large, it is the equivalent of that caused by a very heavy atom. There are severe practical difficulties, particularly as regards the strength of source required, and no structural analysis by Mössbauer gamma rays has yet been made.

The possibility of a gamma-ray laser

It should be possible in principle to observe the stimulated emission of gamma radiation, a process in which an excited nucleus is triggered into the emission of a photon by an incident resonant photon. The question then arises: can the gamma ray equivalent of the maser or laser be constructed? There are two essential ingredients in such a system, a resonant process of stimulated emission which is more probable than other processes, such as non-resonant absorption and inelastic scattering, that effectively remove the resonant photons, and a means of trapping the radiation, for example, in a resonant cavity or between Fabry-Perot plates, in order to increase the probability that a photon will cause stimulated emission rather than simply escape. The second condition could probably be fulfilled by having the resonant nuclei bound in a very perfect single crystal. Such a crystal has the properties of a waveguide for radiation travelling in a narrow angular range close to Bragg diffraction angles, although the efficiency of trapping is not so high as in a Fabry-Perot system. The first condition is fulfilled by ensuring that there is a large density of atoms or nuclei in the upper excited state, so that stimulated emission becomes the most probable process. In atomic systems this is achieved by the 'pumping' process from other excited states. For gamma rays the choice of the 'pumping' radiation is restricted by the availability of nuclear level separations, and the likelihood of matching a pumping source with a suitable resonant system is remote. The alternative would be to populate the upper excited states by radioactive decay from states that were still higher. The difficulties inherent in this method will be clear if we take ^{57}Fe as an example. The cross section for stimulated emission from the 14·4 keV state in ^{57}Fe is about 30 times the photoelectric cross section of iron for this energy. If stimulated emission is to be as probable as photoelectric absorption, then 1 in 30 of the iron nuclei in an iron crystal must be in the excited state. But the excited state has a half-life of only 10^{-7} seconds, so that a crystal containing 10^{22} nuclei (about 1 cm^3) and therefore 3×10^{20} in the excited state will emit approximately 3×10^{27} 14·4 keV gamma rays per second, which corresponds to a power production of 5×10^9 kW. Radioactive transitions feeding into the excited state would correspond to an even greater energy production, much greater in fact than could be dissipated in the crystal without completely vaporizing it.

References

[1] Mössbauer, R. L. *Z. Phys.*, 151, 124, 1958.

[2] Pound, R. V. and Rebka, G. A. *Phys. Rev. Letters*, 4, 337, 1960.

[3] Jaseja, T. S., Javan, A., Murray, J., and Townes, C. H. *Ibid.*, 12, No. 11, A2, 1963.

[4] Champeney, D. C., Isaak, G. R., and Khan, A. M. *Phys. Letters*, 7, 241, 1963.

[5] Hanna, S. S., Heberle, J., Littlejohn, C., Perlow, G. J., Preston, R. S., and Vincent, D. H. *Phys. Rev. Letters*, 4, 177, 1960.

[6] Cordey-Hayes, M. *J. inorg. nucl. Chem.*, 26, 915, 1964.

[7] Boyle, A. J. F., Bunbury, D. St. P., and Edwards, C. *Proc. phys. Soc. Lond.*, 79, 416, 1962.

[8] Delyagin, N. N. and Goldanskii, V. J. 'Proceedings of the Dubna Conference on the Mössbauer effect', 1963.

[9] Cordey-Hayes, M. *J. inorg. nucl. Chem.*, 1964. (In press.)

[10] Flinn, P. A., Ruby, S. L., and Kehl, W. L. Westinghouse Research Laboratories, Paper 63-128-117-P6, 1963.

[11] O'Connor, D. A. and Butt, N. M. *Phys. Letters*, 7, 233, 1963.

[12] Black, P. J. and Moon, P. B. *Nature, Lond.*, 188, 481, 1960.

General references

Frauenfelder, H. 'The Mössbauer Effect.' Benjamin, New York. 1962.

Burcham, W. E. *Sci. Progr.*, 48, 630, 1960.

Boyle, A. J. F. and Hall, H. E. *Rep. Progr. Phys.*, 25, 441, 1962.

The collision, coalescence, and disruption of drops

B. J. Mason

The behaviour of two drops as they approach each other in a dispersion medium is important in rain clouds, in aerosols, and in such processes as distillation and condensation. This article discusses theoretical studies of the relevant factors, and it also describes experimental techniques that have been designed to investigate the problem with the help of photography.

A population of liquid drops, dispersed as an aerosol in a gas or as an emulsion in another liquid, will tend to coagulate to form larger drops. Neighbouring drops tend to collide and coalesce as they move relative to each other under the influence of dynamical, gravitational, or electrical forces; of turbulence; or in the case of very small droplets, of Brownian motion. The relative paths of two approaching drops and hence the probability of their colliding are determined mainly by the hydrodynamical forces set up by the motion of the drops through the fluid; and these forces are governed by the sizes and masses of the drops, their relative velocity, and the density and viscosity of the fluid. If the drops are electrically charged or are subjected to a strong external electric field, their relative motion may also be influenced by electrical forces. In a turbulent fluid, the random fluctuations of velocity may affect neighbouring drops unequally and increase the chance of collision relative to that in a tranquil environment. If two drops collide, they may either coalesce or separate. A barrier to coalescence is presented by the film of fluid trapped between their surfaces, and this has to be expelled through the action of dynamical, gravitational, or electrical forces working against the viscous resistance of the intervening fluid.

The collision and coalescence of drops are matters of widespread importance in, for example, the formation of fog, clouds, and rain; in the stability of aerosols and emulsions; in condensation and distillation processes; and in the thermodynamic properties of droplet-vapour systems. Very similar physical problems arise in connection with the stability of gas bubbles in liquids and are of basic importance in the physics of cavitation and boiling, in the properties of foams, and in the flotation of minerals.

The converse problem, the disintegration of liquid masses into smaller drops by the action of either mechanical, hydrodynamical, gravitational, or electrical forces, is highly relevant to the production of sprays and aerosols: in nature, the break-up of large falling raindrops may, by initiating a chain reaction, accelerate the production of very heavy rain and produce some electrification in thunderstorms.

Collisions between water drops in air

In a cloud composed of water drops of different sizes, a relatively large drop will fall faster than the smaller ones and so may overtake, collide with, and coalesce with some of those lying in its path. As a very small drop—a droplet—approaches a larger drop, it will tend to follow the airflow around the drop and so avoid colliding with it. Alternatively, the droplet's momentum may be too high for it to be deflected completely from its original course and it will then move across the airflow and collide with the drop.

Figure 1 represents the path taken by a droplet of radius r approaching, from afar, a drop of radius R and ultimately just making grazing contact with it. This limiting trajectory of the droplet relative to the drop defines a horizontal circle of radius y_c such that any droplet of radius r whose centre lies initially inside this circle collides with the drop. Thus the cross-sectional area of the cylinder effectively swept out by the drop is πy_c^2, while the maximum geometrical cross-section is $\pi(R+r)^2$. The ratio, $y_c^2/(R+r)^2$, is called the collision efficiency E and has a maximum value of unity.

The actual numerical value of E is dependent upon the sizes and relative velocity of the drop and droplet, and upon the density and viscosity of the air. If the drop is so much larger than the droplet that the latter has negligible influence on the airflow round the former, the hydrodynamical equations can be solved with reasonable accuracy to obtain the relative path of the two particles and hence the collision efficiency. The problem becomes much more difficult when the drop and droplet are of comparable size, so that their airflow patterns mutually interfere. This hitherto intractable problem has been solved by L. M. Hocking [1] for the limited case in which both particles individually obey Stokes's law. For water drops falling in air, this means that the calculations are not accurate for drops of radius greater than 30 μ. The calculated efficiencies with which drops smaller than this will collide with even smaller droplets are shown in figure 2. The inference is that drops of $R < 19$ μ do not capture droplets of any smaller size. Also, for a drop of given size, there is a minimum droplet size below which the collision efficiency approaches zero; thus 30 μ drops fail to collide with droplets of $r < 6$ μ, and 25 μ drops are avoided by droplets of $r < 5$ μ.

The conclusions of these theoretical computations have been tested in my laboratory by R. G. Picknett [2], and by J. D. Woods and B. J. Mason [3]. In one experiment, several thousand equal-sized drops of pure water are allowed to fall freely through a cloud of smaller saline droplets of radius 1–15 μ on to a grease-coated slide. A drop landing on this slide quickly evaporates, but if it has captured one of the saline droplets, it leaves behind a small speck of salt. The slide is now placed in a small chamber that reproduces the temperature and humidity that existed in the original cloud, and is examined under the microscope. Re-condensation of water vapour upon the salt

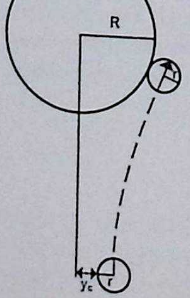


Figure 1 Diagrammatic representation of the trajectory of a small falling drop relative to a large one.

B. J. Mason, *D.Sc.*
Was born in 1923 and educated at University College, Nottingham. In 1947 he was appointed lecturer in Meteorology at the Imperial College of Science and Technology. There he established a research group working on the basic physics of cloud- and rain-forming processes. From 1957 to 1961 he was a Warren Research Fellow of the Royal Society, and was appointed to the first Chair of Cloud Physics at the University of London in 1962.

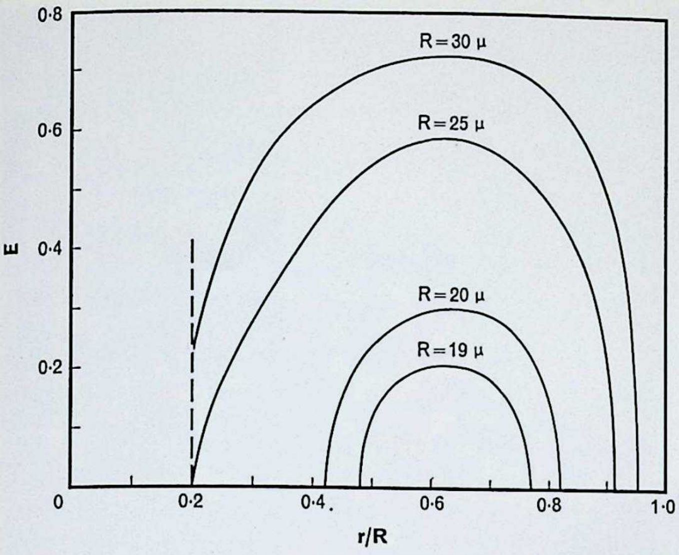


Figure 2 Calculated values of the collision efficiency E for drops of radius R colliding with droplets of radius r.

particles now produces hemispherical drops, the diameter of each being $\sqrt[3]{2}$ times that of the corresponding captured saline droplet. The number of salt droplets in the cloud is low enough for only a negligible fraction of the collector drops to collide with more than one salt droplet. The proportion of collector drops producing a salt particle on the slide is therefore the proportion that has undergone coalescence, and examination of a large number of such drops determines the fraction of these that have coalesced with each size of salt droplet. Having determined the total number and the size of the collector drops introduced into the cloud, and also the sizes and concentrations of the saline cloud droplets by allowing them all to fall out on to another hydrophobic slide, the collection efficiency of drops of this size for smaller droplets of different sizes may be evaluated.

The experiments confirm that drops of radius less than 18 μ are unable to capture droplets of radius 1–15 μ, and the collection efficiencies agree, within the limits of experimental error, with the theoretical values of collision efficiency found by Hocking and others for drops with $R < 60\ \mu$. This agreement strongly suggests that collisions between drops of this size are nearly always followed by coalescence, and that separation after collision is rather rare. This has been confirmed by photographing the interaction of an airborne stream of uniform equally spaced droplets with a similar stream of droplets of rather different size. Figure 3 shows the two streams, the radii of the droplets being 150 μ in one and 120 μ in the other. We see two droplets, one from each stream, in the act of coalescence and, at the bottom of the picture, a drop that resulted from an earlier coalescence.

The airflow round droplets of $R > 30\ \mu$ becomes asymmetrical as a region of unsteady flow, called the wake, develops in the rear. The droplets may then capture an even larger overtaking drop by sucking it into its wake. This may be observed even with two droplets of the same size if they approach each other within a few drop diameters. Figure 5 shows the streaks produced on a film by two illuminated drops of radius 60 μ falling close together. At first we see two separate streaks representing the highlights of the individual drops, but as one drop becomes accelerated in the wake of the other, the two collide and coalesce to form one drop. This drop vibrates violently at first, as shown by the single oscillatory streak, but after five milliseconds the oscillations die away and leave a single straight track. Many such events have been recorded for drops of radius greater than 40 μ. By illuminating the drops with a flash operating at 1000 cycles per second, we can place time markers at intervals of one millisecond on the streaks, and so measure the relative velocities and accelerations of the two drops and compute the collision efficiencies directly from the photographs.

Coalescence of drops

When two drops collide, they may either coalesce or separate again. In an attempt to investigate the physics of the problem, O. W. Jayaratne and B. J. Mason [4] have studied the impaction, bouncing, and coalescence of small water drops at a plane air–water interface.

Streams of droplets, of very uniform size, are produced by the vibrating-capillary device described by Mason, Jayaratne, and Woods [5]. Water is forced at a constant rate through a hypodermic needle attached to the centre of the iron diaphragm of an earphone. By adjusting the frequency of an audio-frequency oscillator connected to the energizing coil of the earphone's electromagnet, the needle may be made to vibrate in a stable resonant mode for long periods. At the turning points in its oscillation, the tip of the needle flings off a drop, all the drops being of identical size. The size and rate of production of the drops are governed by the flow rate of the liquid, the needle diameter, its resonant frequency, and the amplitude of its vibration. By a suitable choice of these parameters, stable streams of very uniform drops may be produced with radii in the range 15–500 μ.

When a stream of uncharged drops falls upon a plane clean surface of freshly distilled water, they may either bounce, as shown in figure 4*a*, or coalesce. Which of these processes occurs depends upon the size of the drops, their velocity, and the angle at which they strike the water surface. The velocities and angles for both the incident and rebounding streams are readily measured from the photographs, because the time interval between successive droplet images is equal to the vibration period of the needle, which may be determined accurately by a stroboscope. Drops striking almost vertically remain in contact with the surface for about one millisecond and lose about 95 per cent of their kinetic energy during impact. This energy is used to form a crater in the water surface and to produce a train of capillary waves that travel out from the point of impact, as shown in figure 6. For nearly normally incident drops, the critical value of the impact angle, θ, at which coalescence occurs decreases as the impact velocity is increased and is smaller for larger drops.

Drops carrying a net electric charge, and drops polarized in an applied electric field, coalesce with the water surface more readily than uncharged drops of the same size and impact velocity. Typically, drops of radius 150 μ impacting with a normal component of velocity of 100 cm/sec coalesce if the charge exceeds 10^{-4} e.s.u., or if the field exceeds 100 V/cm.

The charge on the drops may be gradually increased by applying a rising potential to the vibrating needle and, as the critical charge for coalescence is approached, the rebounding stream becomes faint and consists of smaller droplets than the incident stream (figure 4*b*). The critical charge required to produce this partial coalescence is quite reproducible for drops of given size and impact velocity. Flash photographs reveal that coalescence is accompanied by the ejection of a small jet of water from the surface, and that this jet breaks up and ejects a smaller drop. It is these smaller drops that form the rebounding stream. A further increase in the potential applied to the needle causes complete coalescence, and there is no rebounding stream.

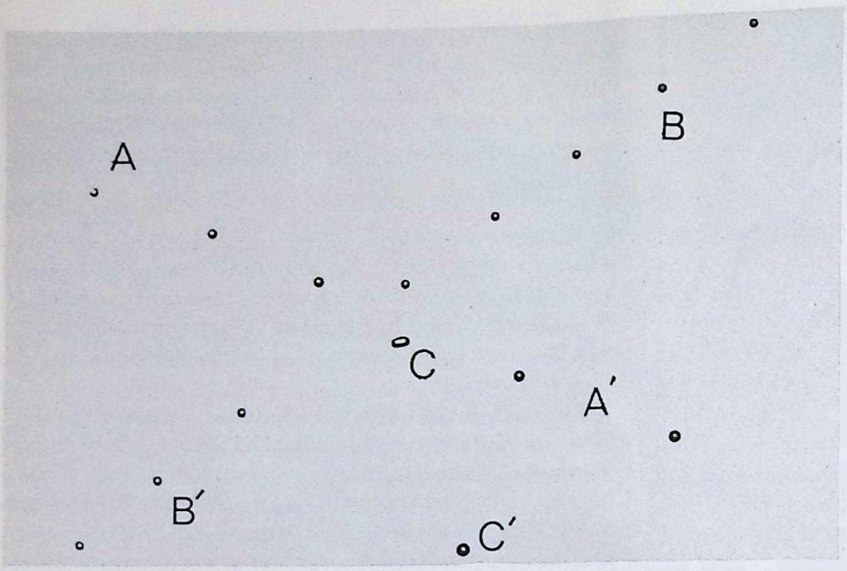
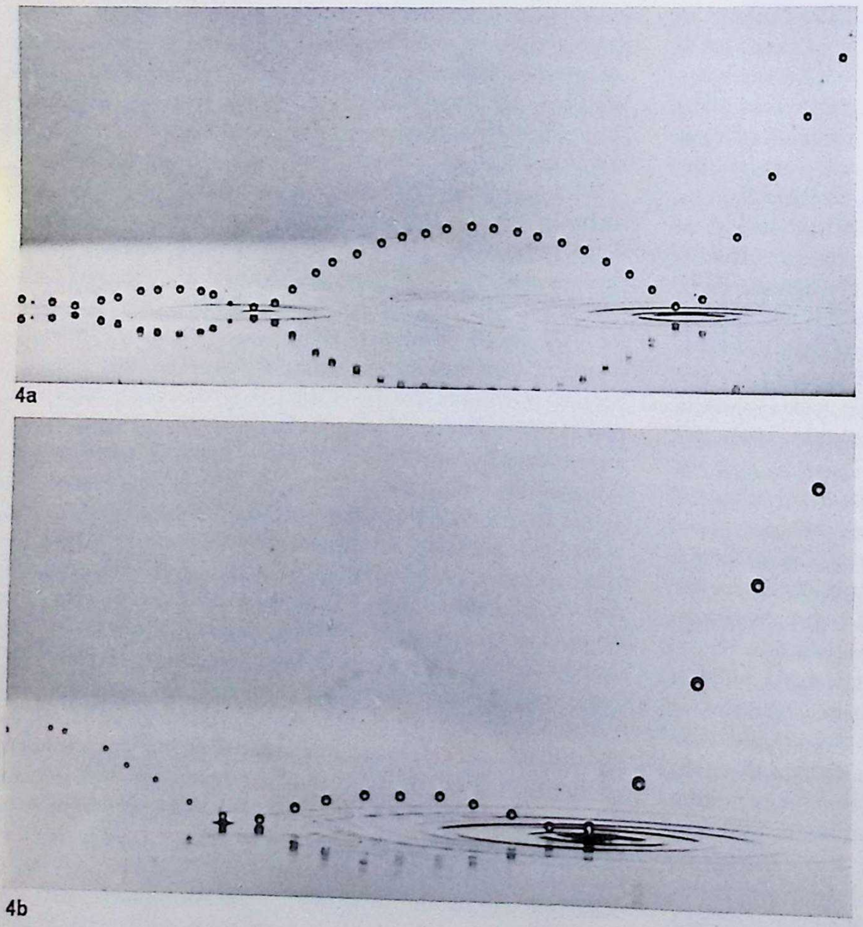


Figure 3 Two streams of droplets AA′ and BB′ interacting to produce larger drops by coalescence at C and C′.


4a

4b

Figure 4 (*a*) A stream of drops of radius 150 μ undergoing three successive stages of bouncing at a plane water surface. The changes in velocity of the drops are indicated by changes in their spacing. (*b*) Two stages of partial coalescence between a stream of water drops and a plane water surface. In each case the reflected stream is composed of smaller drops than the incident stream.

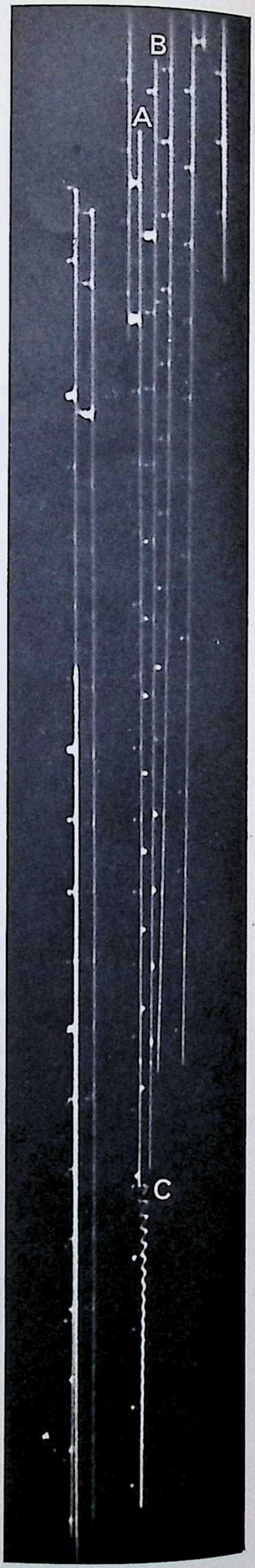


Figure 5

Figure 5 A and B are the highlights of two equal-sized drops, one falling in the wake of the other. At C, they coalesce to form one drop that vibrates and produces the single oscillatory streak.

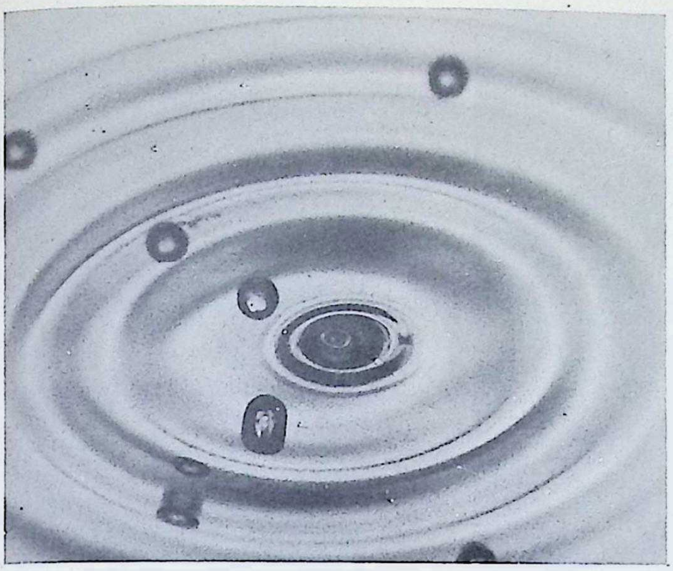


Figure 6 Drops of radius 150 μ impacting at 100 cm/sec on a plane water surface to produce a crater and train of ripples. The upper surface of the drop develops a dimple during impact.

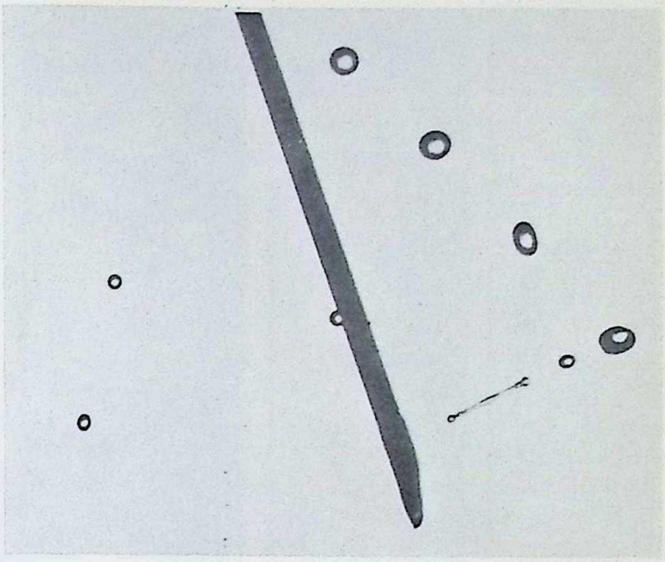
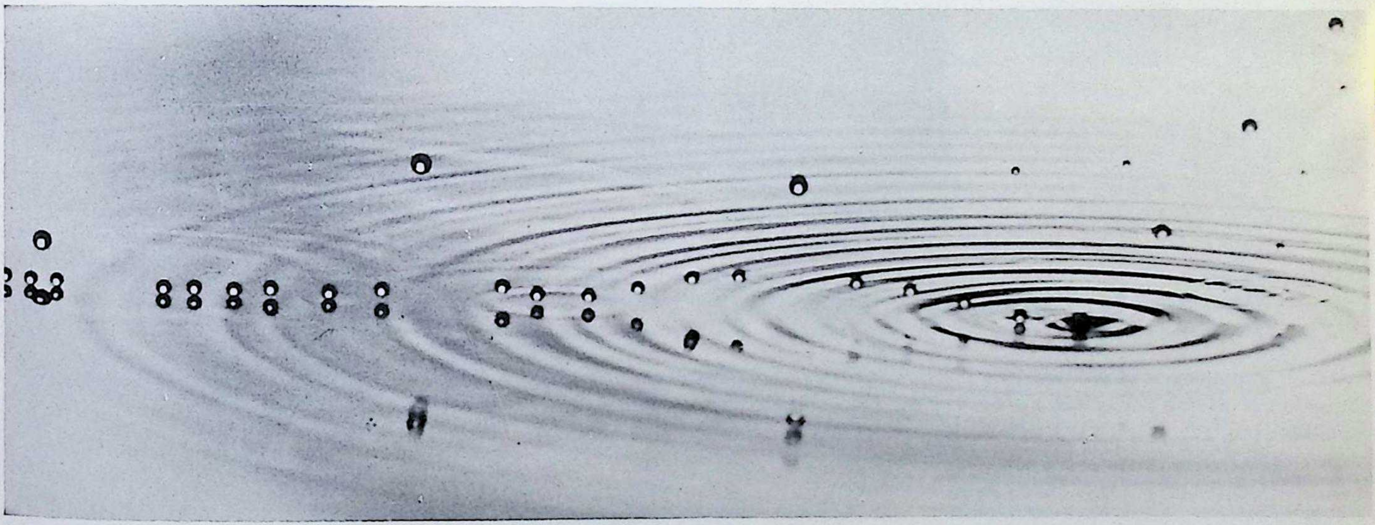
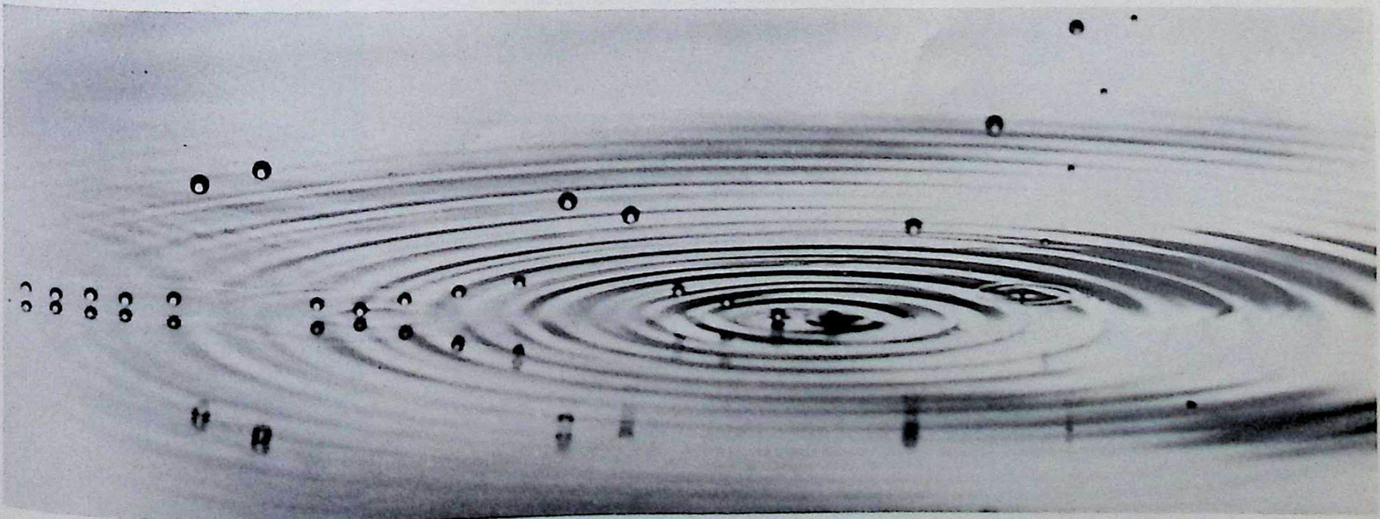


Figure 7 The break-up of a thread of liquid issuing from the tip of a vibrating hypodermic needle.

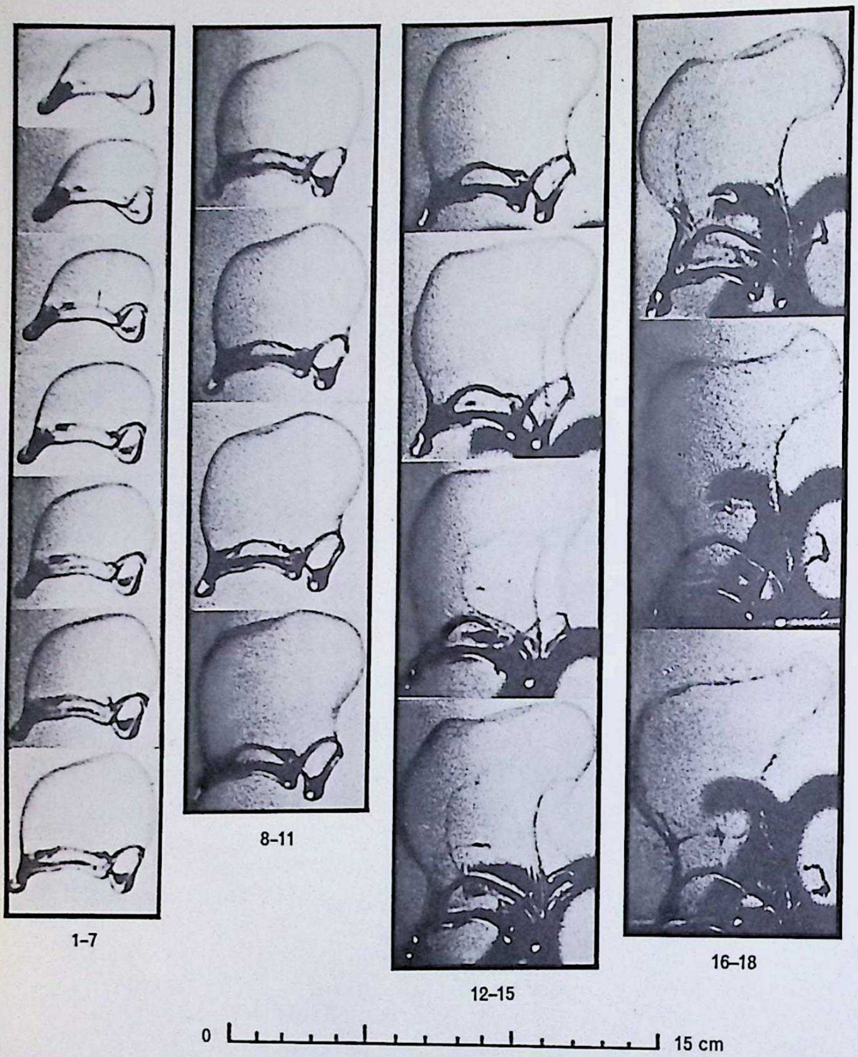


8a



8b

Figure 8 (*a*) Partial coalescence of a stream of water drops is induced by applying a steady potential of +50 V to the hypodermic needle. But this potential is neutralized periodically by −50 V pulses of duration 2·5 msec, so that every seventh drop bounces. (*b*) As for (*a*), but the pulse duration is increased to 4·5 msec, so that the bouncing drops occur in groups of two.

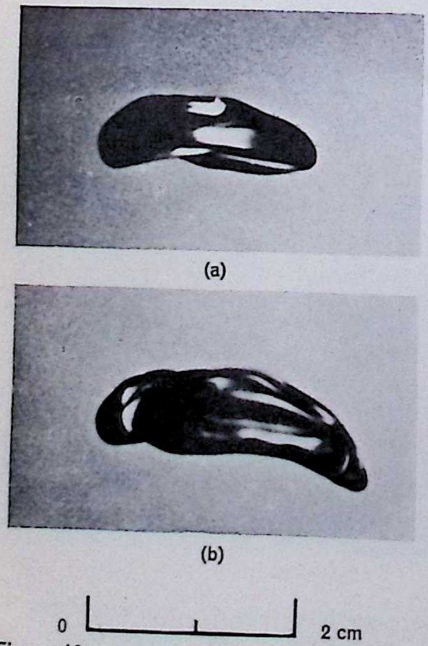

1–7

8–11

12–15

16–18

Figure 9 A sequence of photographs taken at intervals of about 1 msec showing the disruption of a large falling water drop. The drop blows up to form a rapidly expanding bag supported on an annular ring of liquid which breaks up into several large drops.


(a)

(b)

Figure 10

Figure 10 A large falling water drop flattens and develops a depression in its base before breaking up.

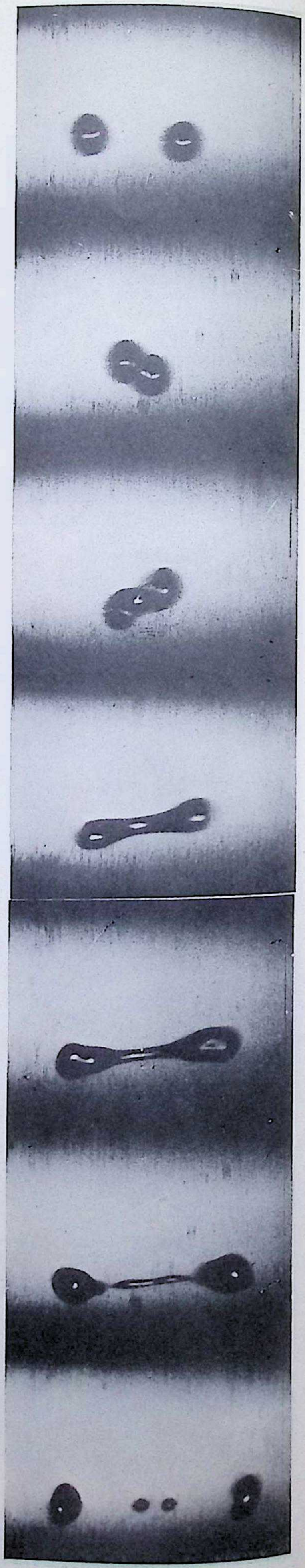


Figure 11

Figure 11 The coalescence and subsequent break-up of two water drops of radius 130 µ colliding at a relative velocity of 3 m/sec. Photographs taken at 7000 frames per sec.

If, in addition to the steady potential, the needle receives voltage pulses of equal magnitude but opposite polarity, the drops can be made to bounce during the duration of the pulse but to coalesce with the water surface in the intervals between pulses. The frequency and spacing of the drops in the reflected stream can therefore be controlled by varying the duration and repetition rate of the square-wave pulses (figure 8).

A simplified theory for the motion of a drop rebounding from the water surface yields expressions for the depth of the crater and the restoring force at any stage, and also for the time of contact and the energy lost during impact. The computed values of these parameters are in reasonable accord with experiment. If the drop is to coalesce with the water surface, it must first expel and rupture the intervening air film.

If the film has to reach a certain minimum thickness before coalescence can occur, the theory predicts that the greater the radius and impact velocity of the drop, the more difficult will coalescence be, because the film is less likely to attain this minimum thickness during the period of contact. These predictions are confirmed by the observations, but there are departures for drops striking at nearly glancing incidence and for large energetic drops striking nearly normally.

Drops impinging nearly tangentially have little penetrating power, produce only shallow craters, and tend to skim along the surface, continually renewing the air film. Coalescence is thereby impeded, and these drops remain on the surface for about 100 msec, or 100 times longer than similar-sized drops impinging at nearly normal incidence.

Relatively large drops, of radius greater than about 150 μ, impinging nearly normally on the water surface, are considerably deformed during impact. This is demonstrated in figure 6, which shows that a distinct dimple appears in the upper surface of the drop, and it seems very likely that the lower surface would follow suit. The effective area of the trapped air film may now be confined to a narrow annular ring, making for more rapid draining of the film and easier coalescence. Since the deformation is greater for larger drops striking at higher velocities, it follows that larger drops should have smaller critical velocities for coalescence. This is observed to be so.

The fact that a charge or a polarizing electric field always aids coalescence is a little surprising because, for the charges and fields used in the experiments, the electrical forces of attraction between the drops and the earthed water surface, even if these are separated by as little as 0·01 μ, are calculated to be very small compared with the dynamical forces produced by deceleration of the impacting drop. It may be that the electrical forces distort the liquid surfaces so that drainage of the air film and the formation of a liquid bridge are facilitated. An undistorted drop of radius 150 μ, polarized in an external field of 100 V/cm and approaching within 0·15 μ of the plane surface, would produce a field of 40 kV/cm in the air gap. Such a field should cause the drop to elongate in the direction of the field and might rupture the air film by producing a micro-discharge.

The disruption of drops

The disintegration of bulk liquid into small drops is usually achieved by forcing the liquid under pressure through a small hole. The liquid issues in the form of a thin unstable thread which becomes varicose and breaks up into a series of drops under the action of surface tension. The phenomenon was first investigated by Lord Rayleigh [6], who showed that a jet of circular cross-section is unstable if the wavelength of the disturbance exceeds the circumference of the undisturbed jet. As an example of this, figure 7 shows a thread of water issuing from the tip of a vibrating hypodermic needle and breaking up into drops of different sizes. In this case, the continuously vibrating needle produces separate streams of drops, the drops in any one stream being very uniform in size.

As we have seen, when two small water drops, of radius less than 100 μ, collide in air with relative velocities of only a few tens of centimetres per second, they usually coalesce and form a single stable drop. If, however, two such drops are projected towards each other at relative speeds of a few metres per second, they may well coalesce, but the resulting drop oscillates very violently and may break up into several fragments. This is shown in figure 11.

Very large water drops, greater than about 6 mm in diameter, break up spontaneously when falling at terminal velocity in air. This is why raindrops only rarely exceed this size. The shape and terminal velocity of the drops falling in air are governed by the aerodynamic, hydrostatic, and surface-tension forces which, acting in concert, determine the pressure distribution over the surface of the drop. Gross distortion and even disruption of the drop occur if the aerodynamic forces overwhelm the others. The break-up of such large drops has been photographed by a number of workers [7–10].

Under the dominant influence of the aerodynamic forces, the drop progressively flattens and assumes a lenticular shape (figure 10). The base then develops a concave depression and, soon after, there develops a rapidly expanding bubble or bag supported on an annular ring containing the bulk of the liquid. The final stage of the break-up occurs when the bag bursts and throws off a fine spray of droplets, while the annular ring breaks up into several large fragments. These events are shown by the photographs of figure 9 taken at millisecond intervals.

In normal conditions, the large fragments acquire a positive charge and the fine spray from the bursting bag a compensating negative charge. This was first discovered by G. C. Simpson [11], who showed that the charge was about 0·02 e.s.u. per gram of water. Simpson thought that the break-up of large raindrops was the main mechanism of charge generation in a thunderstorm but, as Mason [12] has argued recently in this journal, it is much too weak and, in any case, produces charges of the wrong polarity. However, Matthews and Mason [10] have recently studied the break-up of large drops in the presence of vertical electric fields of up to 2 kV/cm such as might exist in a mature thunderstorm. The charges on the large fragments are now one hundred times larger than those produced in the absence of a field and those measured by Simpson, so it appears that breaking drops may be responsible, not for the primary charging mechanism of the thunderstorm, but for the subsidiary positive charge that often appears near its base.

References

[1] Hocking, L. M. *Quart. J. R. met. Soc.*, 85, 44, 1959.
[2] Picknett, R. G. 'Aerodynamic Capture of Particles', p. 160. Pergamon Press, London. 1960.
[3] Woods, J. D. and Mason, B. J. *Quart. J. R. met. Soc.* (In press.)
[4] Jayaratne, O. W. and Mason, B. J. *Proc. roy. Soc. Lond.* (In press.)
[5] Mason, B. J., Jayaratne, O. W., and Woods, J. D. *J. sci. Instr.*, 40, 247, 1963.
[6] Rayleigh, Lord. *Proc. roy. Soc., Lond.*, 29, 71, 1879.
[7] Hochschwender, E. Dissertation, Heidelberg University. 1919.
[8] Lane, W. R. *Industr. Engng Chem.*, 43, 1312, 1951.
[9] Magarvey, R. H. and Taylor, B. W. *J. appl. Phys.*, 27, 1189, 1956.
[10] Matthews, J. B. and Mason, B. J. *Quart. J. R. met. Soc.* (In press.)
[11] Simpson, G. C. *Phil. Trans.*, *A*, 209, 379, 1909.
[12] Mason, B. J. *Endeavour*, 21, 156, 1962.

Physical measurement by gas chromatography

J. H. Purnell

Although gas chromatography is very widely used for analysis, few are aware of its potential value for physical measurement. Studies of column performance can yield information about solubility, adsorption, thermodynamic quantities, and diffusion. Specialized techniques permit measurement of solid surface areas, boiling points and vapour pressures, and, with developing understanding, the assignment of molecular structure. One of the attractive features of this particular technique is its technical simplicity and the rapidity with which information is obtained. The field is only just beginning to be explored, but already it shows great promise.

Although gas chromatography is widely used for analysis, and its power and breadth of application in this field are very well recognized, few are aware of its potential as a purely physico-chemical technique. Broadly, two quite different types of study are possible. Firstly, measurement of the residence times of substances in a column allows determination of partition and activity coefficients. By studying the effects of changing the experimental conditions, partition or adsorption isotherms and the relevant thermodynamic quantities can be evaluated. As there is growing evidence of a relationship between residence time and structure, it appears likely, also, that the method may become a useful ancillary in structural investigations. Secondly, observation of the rate of broadening of individual bands of material as they pass through a column permits calculation of gas phase interdiffusion coefficients, and eventually, it is hoped, of liquid-phase interdiffusion coefficients.

The attraction of making such studies by gas chromatography is the relative simplicity of the apparatus and, in comparison with most conventional methods, the smaller demands on experimental skill. Further, experiments can be conducted very quickly, and consequently the scale of investigations may be extended well beyond what would be regarded as reasonable with other methods.

While the areas of study outlined above do not exhaust the possibilities and, indeed, several others have been suggested and in some cases tested, the major effort so far has been along these lines. This article outlines the basis of such studies and indicates, so far as is at present possible, the quality of the results obtained.

Experimental

Figure 1 illustrates the basic construction of a gas chromatograph. The heart of the apparatus is, of course, the column, and this may be of either the packed or the capillary type. Packed columns are most often made of stainless steel tube of 0·3–0·6 cm diameter. For compactness, they can be coiled, straight, or zigzagged. The packing may be either a granular solid adsorbent such as silica, alumina, or molecular sieve—the last for the analysis of low molecular weight gases—or a low area (*ca.* 0·5–4 m^2/g) porous, solid support coated with an involatile liquid. Capillary columns are normally of about 0·025 cm internal diameter and are coated internally with finely powdered adsorbent or a thin liquid film.

Carrier gas, for example H_2, He, N_2, or Ar, is drawn from a cylinder and passes first through a purifying train designed primarily to remove water and the vapours of oils and greases. After passage through metering valves, it flows successively through an injection system, the column, a detector, and (usually) a flowmeter, and finally escapes to the atmosphere. A thermostat is needed, as the performance of the column is very much dependent upon the temperature and its accurate control. For analytical purposes, control to about ±0·5°C is probably adequate. For physical measurement, however, very much better control is mandatory, and it is unlikely that commercial equipment is as yet equal to the requirements.

Many different detectors are now available. Much the most widely used, however, are the thermal conductance cell (katharometer); the hydrogen-flame ionization detector; and one or other of the variants of the electron-capture radiological detector. When the last of these is used with argon as carrier gas, electrons, usually from a tritium source, produce metastable argon which very efficiently ionizes organic molecules in collision. When other carriers are used, the detector is of much lower sensitivity, since it then relies on the more effective capture of electrons by organic than by carrier molecules. Almost all detectors are to some extent sensitive to flow fluctuation, even of pure carrier gas, and it is common practice to seek to eliminate this by running detectors in pairs. As indicated in figure 1, the second (reference) gas flow may be drawn off ahead of the injection system, although other arrangements are possible. Many workers choose to incorporate a second column within the thermostat in this reference stream, because this adds flexibility and provides a convenient and readily controlled resistance to flow. Detectors may also be

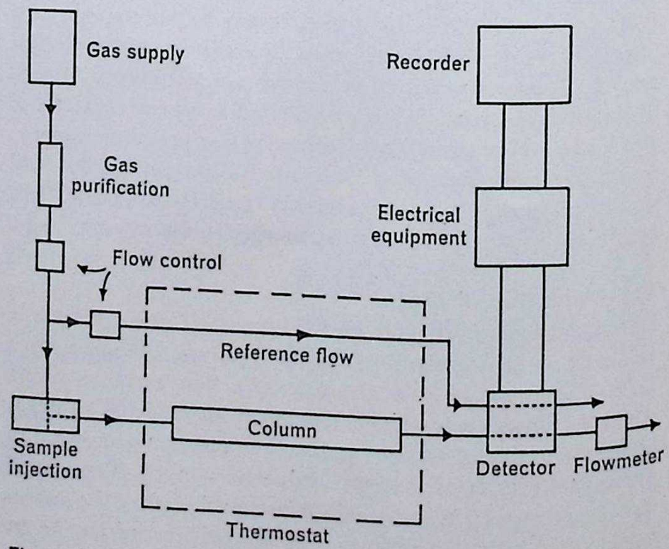


Figure 1 Block diagram of gas chromatograph. Arrowed lines show direction of gas-flow.

J. H. Purnell, *M.A., B.Sc., Ph.D.*

Was born in the Rhondda Valley, South Wales, in 1925 and educated at Pentre Grammar School; University College, Cardiff; and the University of Cambridge. He was Lecturer in Chemistry at Cardiff *(1947-52)*, Fellow of University of Wales *(1953–55)*, and Demonstrator in Physical Chemistry at the University of Cambridge *(1955–60)*, where he is now Lecturer in Physical Chemistry. His chief research interests are gas kinetics and chromatography.

sensitive to ambient temperature changes; this is particularly true of katharometers, and the double-channel system again provides compensation.

In practice, the desired rate of flow of pure carrier gas is first established and any signal transmitted from the detector to the recorder is balanced out in the ancillary electrical equipment. This provides a base line which is nominally the zero of detector output. A sample is then introduced into the system, and the emergence of each of its components is marked by an off-balance signal from the detector to the recorder resulting from the contamination of the carrier gas. Every effort is made to adjust conditions so that the detector signal is linearly proportional to the amount of contaminant.

Chromatographic techniques

There are two important basic ways of using a chromatographic system.

Elution chromatography. In this, a discrete sample of material is introduced into the carrier-gas stream. During passage through the column, a given component is distributed over and over again in constant ratio between the gas and immobile phases. This ratio is governed by a fundamental physical quantity, the partition coefficient. If the various partition coefficients differ sufficiently, each component of a mixture emerges from the column as a separate band. In fixed conditions of temperature and flow rate, the time of emergence of a band is characteristic of the system. The repeated distribution of material between phases gives a more or less Gaussian concentration profile along the column for each component, and this gives rise to the familiar elution chromatogram comprising a set of bell-shaped peaks (figure 2).

Frontal development. This is carried out by injecting into the carrier-gas stream a continuous supply of the sample material. When the sample front enters the column there is again the recurring distribution between phases, but, because the depletion of the gas phase is constantly made up by incoming sample, peaks do not develop. Instead, the concentration profile takes the form of a plateau preceded by an S-shaped front. In the region of a plateau the immobile phase is saturated with respect to that components and fronts 'break through' the column only when the whole of it is saturated. The breakthrough times are again characteristic of the system, and the chromatogram takes the form of a series of steps (figure 3). After the last component has emerged, the column serves only to delay the flow, and the input and output compositions are identical.

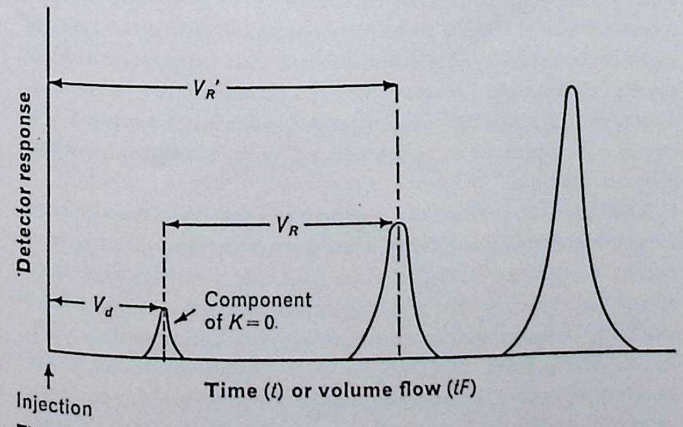


Figure 2 Elution chromatogram. Abscissae can be plotted in units of time (t) or of integrated volume flow (tF); F is rate of flow of gas. The first peak represents an unsorbed component; its retention volume thus measures the total free gas space (dead volume) in the apparatus between injector and detector.

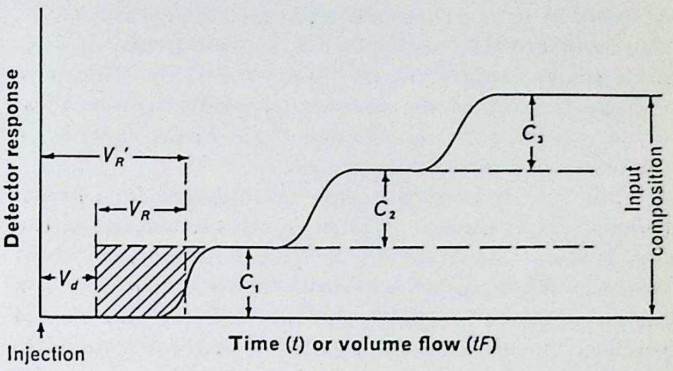


Figure 3 Frontal chromatogram of three-component mixture. The shaded area represents amount of component 1 absorbed by column packing; step heights are proportional to individual concentrations (c).

Retention theory

Each of the numerous approaches to chromatographic theory [1–3] leads to a very simple relation between the total volume of carrier gas that flows through the column in the time between injection and emergence of a given elution band (the apparent retention volume, V_R'); the free gas volume V_d, the volume of immobile phase, V_s; and the partition coefficient K. The retention equation

$$V_R' = V_d + KV_s \qquad \ldots.(1)$$

is strictly valid only for the condition of zero pressure drop across the column, because a finite pressure gradient leads, through decompression of the gas, to an artificial expansion of V_d and KV_s. However, the correction procedure is established [1] and simple to apply; in further discussion it is assumed to have been used.

The quantity V_d is of no theoretical significance, since it comprises only the total free gas volume between the point of injection and the detector and so depends entirely on the geometry of the system. Thus a 'true' retention equation may be written

$$V_R = KV_s \qquad \ldots.(2)$$

from which, by application of Henry's law [4], it is a simple matter to show that

$$V_R = \frac{RTw}{\gamma p^0 M_s} \qquad \ldots.(3)$$

R is the gas constant; w and M_s are respectively the weight and molecular weight of the immobile phase; p^0 is the saturation vapour pressure of the particular sample component at the column temperature T; and γ is the mean activity coefficient of the sample component in the immobile phase. V_R may be recalculated to its value at 273°K per gram of immobile phase (specific retention volume, V_g) when, if the units used are ml and mm Hg,

$$V_g = \frac{1{\cdot}7 \times 10^7}{\gamma p^0 M_s} \qquad \ldots.(4)$$

This equation is of surprising simplicity, and because V_g, p^0, and M_s are all readily measurable, opens up the route to study of solution and adsorption processes through measurement of γ. Because the concentration of sample in solution diminishes continuously along the column during an elution, worthwhile use of equation 4 demands experiments with minute samples so that the condition of infinite dilution is approximated even at the injection end of the column. Thus activity coefficients measured by elution chromatography are all, in principle, γ^∞. This restriction does not, of course, apply in frontal development studies. Indeed, in this case γ^∞ may be difficult to measure because of the problem of producing mixed gas streams at sufficiently high dilution. Thus, the two methods are complementary rather than competitive.

While activity coefficients may be calculated from frontal chromatograms through equation 4, the chromatograms can also be used in another way to provide information about partition or adsorption isotherms. In the simple case of a single-component sample stream, which is the only case of practical interest here, the amount of material required to saturate the column is clearly that abstracted from the carrier before break-through. Diagrammatically, as shown in figure 3, this is the amount contained in the shaded area. The volume sorbed, V_a, is then related to the retention volume of carrier gas, V_R (measured from an unsorbed elution peak injected simultaneously with the sample) to the mid-point of the sample front by

$$V_a = V_R\left(\frac{\bar{p}}{p_c}\right) \qquad \ldots.(5)$$

Here $(\bar{p}/p_c)$ is the ratio of sample to carrier-gas partial pressures at the mean column pressure, and also at the inlet as well as at the outlet after break-through. In order that the sample partial (equilibrium) pressure does not change significantly along the column as a result of expansion, a trivial pressure gradient is desirable in practice.

By repeated experiments with varying ratios of $(\bar{p}/p_c)$ a series of values of V_a/w and $\bar{p}$ are obtained and, hence, the sorption isotherm. The main limitation to which the method is subject with simple apparatus is the difficulty of producing streams diluted more than about 1000 : 1; even so, this readily permits work with $\bar{p}$ as low as 1 mm Hg. With elaborate apparatus, multiple stream mixing and hence much greater dilutions can be dealt with. Several studies of adsorption have been carried out in this way: for example, the adsorption of a number of common organic vapours by active charcoal [5] and of acetone and benzene by low-area silicas [6, 7]. In practice, the method offers certain advantages over established techniques because it can be used for the study of very weak adsorbents (for example, surface area < 1 m²/g) and provides the information rapidly.

An elegant alternative approach which combines both elution and frontal methods has recently been suggested by C. N. Reilley [8] and is currently being used by Conder and the author in an attempt to measure the adsorption of such substances as the lower paraffins, olefines, and metal alkyls on glass and quartz. In this, the column is operated in the normal frontal manner until the plateau breaks through. Then, discrete and minute volumes of either the sample or of the carrier are injected and eluted; the retention volumes are thus measured at known concentration in the immobile phase. Any carrier gas injected travels through the column as a 'negative' sample. The injections may be repeated as often as wished at any value of $\bar{p}$. At the conclusion of the experiment, a desorption may be carried out by suddenly stopping the sample supply and replacing it with carrier. An illustrative chromatogram is shown in figure 4. In practice, the plateau, after break-through, is biased out electrically to make possible the increase in detector sensitivity that is needed to record the small elution peaks. Alternatively, the detector signals may be differentiated.

There is no doubt that this combination method is very satisfactory and may ultimately supplant the individual methods entirely.

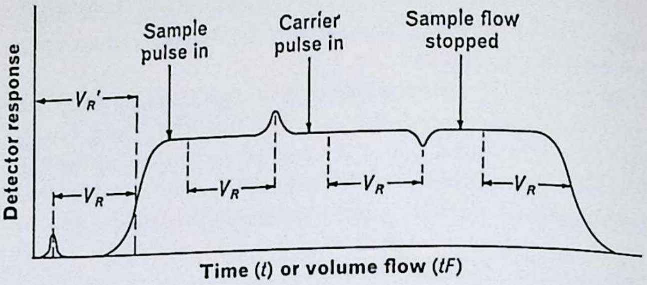


Figure 4 Combined elution/frontal chromatogram. The first peak represents an unsorbed component injected simultaneously with sample front.

Experimental studies

The literature already contains a number of papers devoted to the measurement of γ^∞ by the elution method. It is only very recently, however, that the important question of whether or not gas chromatographically measured data agree with those measured conventionally has been answered with certainty [9–13]. It now appears that, provided γ^∞ is within about an order of magnitude of unity, closely comparable results are obtained. In illustration of this point, table 1 lists some data for $\log_{10}\gamma^\infty$ for a number of solutes in dinonylphthalate at 30°C, measured both conventionally and chromatographically by D. H. Everett and his colleagues [9, 10].

Table I

Values of $\log_{10}\gamma^\infty$ in solutions of dinonylphthalate at 30°C

Solute	$\log_{10}\gamma^\infty$	
	By gas-liquid chromatography [10]	By conventional method, extrapolated [9]
Cyclohexane	$\bar{1}\cdot968$	$\bar{1}\cdot968$
Benzene	$\bar{1}\cdot747$	$\bar{1}\cdot747$
n-Pentane	0·034	0·035
n-Hexane	0·077	0·076
n-Heptane	0·117	0·114
n-Octane	0·153	0·155

The agreement is seen to be excellent. Experience indicates that γ^∞ may quite readily be measured to ± 1 per cent with some certainty. With some sophistication of apparatus and technique, considerable improvement can unquestionably be made. This would be worth while, because even ± 1 per cent accuracy may not be good enough for accurate evaluation of excess enthalpies, as these are normally very small in chromatographic systems.

The need to restrict the method to systems of moderate γ is not definitely established; what doubts there are spring from recent studies by R. L. Martin [14]. He has interpreted his results as showing that the properties of a solvent spread out as thinly as is normal in chromatographic columns may differ considerably from those of the bulk liquid. That this might occur is, of course, reasonable; the Gibbs isotherm, for example, suggests discrepancies of about the size observed by Martin. However, in each case where anomalies were seen, γ^∞ was in excess of 10, and in some instances exceeded 200. It is not yet certain that this problem is one of general significance. If it is, it somewhat complicates solution studies, but, at the same time, it may offer an approach to an area of study which has hitherto been regarded as very difficult.

The agreement between V_g (and hence γ^{∞}) as measured in different laboratories may be judged from the data in Table II, which have been quoted by M. R. Burnett [15].

Table II

Comparison of V_g determinations by different workers. Solvent, squalane; temperature, 80°C

Solute	Specific retention volume, V_g (ml)			
	Ref. [11]	Ref. [15]	Ref. [16]	Ref. [17]
n-Pentane	25·1	27·3	24·7	24·7
n-Hexane	60·7	64·1	62·5	61·7
n-Heptane	145	148	145	144
n-Octane	339	334	330	334

The data clearly indicate the degree of reproducibility that may be achieved, particularly for substances of large V_g. Other instances of such agreement might be quoted for quite different types of solute/solvent system. When account is taken of such factors as the varying origin of the solvent, the probable difference in sample size used by different workers, and the fact that the data are obtained variously with packed and capillary columns (and in one instance [15] by an entrainment method), the results must be regarded as encouraging.

One of the interesting features of the results obtained up to now is that most typical gas-liquid chromatography (GLC) systems, which comprise small solute and large solvent molecules, deviate from ideality of solution mainly through their entropies rather than through their enthalpies of mixing. In some respects, therefore, such systems may be more amenable to theoretical treatment than apparently simpler systems would be. This is likely to stimulate theoretical study and papers along these lines have already appeared, for example [18]. Further, in at least one instance, evidence has been obtained for a linear relationship between partial molar excess enthalpies and entropies [19]; some results are illustrated in figure 5. This finding is of interest and, in other fields, has been a matter of discussion for a considerable period: see, for example, [20].

Frontal development studies have so far been restricted to gas-solid adsorption systems, but we may expect development in solution studies in the future. Figure 6 illustrates one of a number of adsorption isotherms obtained [7] by the frontal method for adsorption of benzene vapour by a hexamethyldisilazane-treated diatomaceous earth having a surface area of only about 0·8 m^2/g. The extent of adsorption is seen to be close to one micromole per gram of solid, a value so low that it would be very difficult to determine by any other method. This underlines the further utility of chromatography in the measurement of small surface areas.

The types of study described so far have implicitly involved the use of a nonvolatile immobile phase. By suitable modification, volatile solvent systems also can be studied. M. R. Burnett, for example, has described [15] a method in which solvent and solute, contained in a special vessel ahead of a column, are entrained away by a constant flow gas stream, and intermittent GLC analysis is used to measure gas-phase concentrations. The method is simple, does not require absolute detector calibration, and is well suited, for example, to the study of aqueous solutions. It was, in fact, developed to facilitate the analysis of food extracts, which commonly consist largely of water. Data obtained for aqueous solutions of ketones and alcohols compare well with those measured conventionally.

Structural and physical properties

Equation 4 is capable of further development. For example, since the temperature dependence of V_g is that of $(\gamma p^0)^{-1}$, for a given sample component, the simplest integrated form of the Clausius-Clapeyron equation leads to

$$\ln V_g = \frac{\Delta H_s}{RT} + k_1 \qquad \ldots.(6)$$

where ΔH_s is the molar heat of evaporation of the component from solution in the immobile phase. Further, because in many instances $\Delta H_s \simeq \Delta H_v$ (that is, the molar heat of vaporization of the pure substance), then from Trouton's rule:

$$\log_{10} V_g \simeq \frac{5T_b}{T} + k_2 \qquad \ldots.(7)$$

where T_b is the relevant boiling point. Finally, since it is well known that for many homologous or otherwise related series

$$T_b = k_3 n + k_4$$

where n is, for example, the number of carbon atoms in a compound,

$$\log_{10} V_g = \frac{k_5 n}{T} + k_6 \qquad \ldots.(8)$$

The general validity of equation 6 has been demonstrated many times, and exceptions are rare. In many instances, the partial excess enthalpy of mixing is so small that graphical measurement of ΔH_s, following elution at several temperatures, provides a good estimate of ΔH_v. This method is of most value for substances having unusual or large molecules, especially since these often cannot readily be prepared in a pure state or obtained in sufficient quantity to permit direct measurement of the heat of vaporization.

Equations 7 and 8 also have been found to apply widely. Figure 7 illustrates some data, plotted after equation 7, for a wide range of mainly aliphatic hydrocarbons (C_5–C_7) eluted from a squalane column at 43°C. It is seen that in this case the points fall on a straight line irrespective of chemical type. Other immobile phases distinguish between types, but still yield linear plots within each homologous series. From such plots, unknown boiling points may be estimated with considerable accuracy. By comparison with saturation vapour pressure data for even a single homologue, reasonably accurate p^0 data may be obtained by elution at various temperatures. For example, p^0 for n-butane at 21°C has been determined chromatographically [4] as 1642 mm Hg, which compares well with the conventionally determined value of 1634 mm Hg; the boiling point was estimated correctly to within 0·5°C.

Equations 7 and 8 also open up a new area as yet almost completely unexploited, namely that of structural assignment. In the simplest case, this might be based entirely on equation 8. For example, it is seen from table I that there is a constant increment of about 0·04 log units in $\log_{10}\gamma^{\infty}$ for each added —CH$_2$— group in the normal paraffin series. A similar, and much larger, increment in log V_g occurs. Table II shows that each additional —CH$_2$— gives an increase of 0·37 units in log V_g. In at least one system [21], the constancy of the increment up to C_{35} has been demonstrated. The regularities that have been observed are not, however, restricted to such simple structural changes. Introduction of alkyl or other groups in different positions in the side chains, unsaturated linkages in various positions, and *cis-trans* isomerism, for example, are usually accompanied by characteristic increments. An early example of the use of such information for structural assignment is found in the work of G. H. Beavan and his colleagues [22,

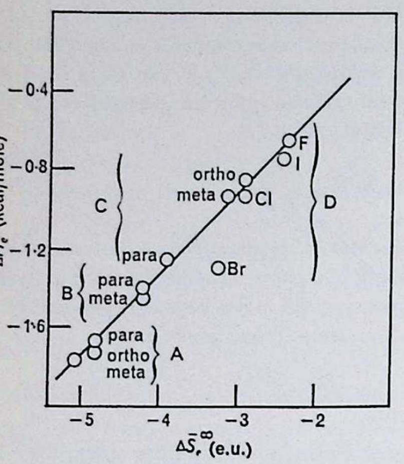


Figure 5 Plot of chromatographically measured partial excess molar enthalpies and entropies at infinite dilution for halogenated aromatics dissolved in 7,8-benzquinoline at 100°C. A, dichlorobenzenes; B, chlorobromobenzenes; C, chlorotoluenes; D, halobenzenes [19].

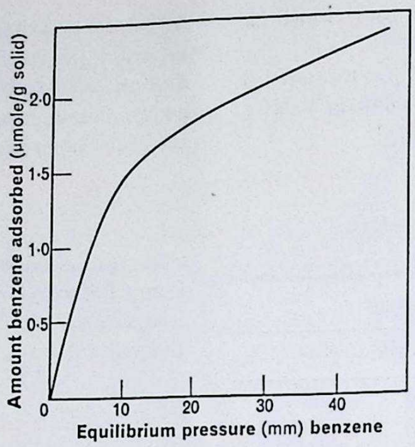


Figure 6 Chromatographically determined isotherm for adsorption of benzene vapour at 50°C by hexamethyldisilazane treated diatomaceous earth (B.E.T. measured area ≃ 0·8 m²/g) [7].

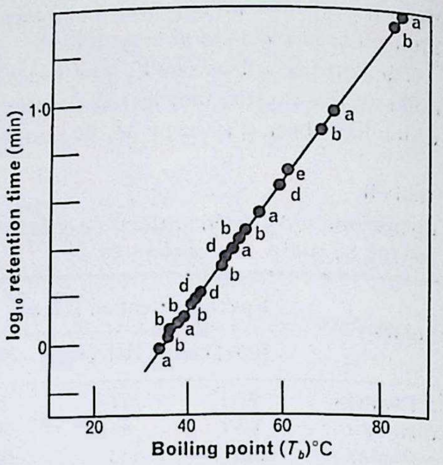


Figure 7 Log of retention time plotted against boiling-point for elution of C_5–C_7 hydrocarbons from squalane column at 43°C. (*a*) paraffins, (*b*) olefines, (*c*) dienes, (*d*) alkynes, (*e*) benzene.

23]. In a detailed study of the retention behaviour of twenty-four alkyl substituted diphenyls eluted from Apiezon-M columns, they found that, independently of the identity of the alkyl group, *meta*-substitution yielded a V_g increment (i) (relative to V_g for diphenyl) per side chain carbon atom of $1{\cdot}5 \pm 0{\cdot}05$; *para*-substitution gave $i = 1{\cdot}6 \pm 0{\cdot}05$, and *ortho*-substitution—that is, in the 2 or 6 position—gave $i \leqslant 1{\cdot}25$. In poly-substituted compounds, the presence of a group in the 2 or 6 position entirely determined the value of i. On the basis of these results, it proved possible to identify several hitherto unassigned alkylated diphenyls. A particularly interesting feature of this work was the excellent correlation of the chromatographic shift data with those obtained by study of ultra-violet spectra. This, if nothing else, emphasizes the fundamental nature of the chromatographic information.

In a very recent study of the retention behaviour of a number of dialkylethylenes whose configuration has been established by stereospecific synthesis, infra-red spectra, and other conventional means, R. A. Hively [24] found that the retention volume ratio of *cis* to *trans* compounds (R) was always greater with a polar immobile phase than with a non-polar immobile phase. In addition, in each case, the sequence of boiling points was $T_b(trans) > T_b(cis)$. The properties of the *cis*/*trans* isomers of 3-methyl-2-pentene and 3,4-dimethyl-2-pentene, which have recently been unambiguously assigned [30] (during which process the names were inverted), were also in accordance with these rules. Using β,β′-oxydipropionitrile as the polar solvent and squalane as the non-polar solvent, Hively found $R_{polar}/R_{non\text{-}polar} \simeq 1{\cdot}08$. In the study of the retention behaviour of 3-methyl-2-hexene and 3-methyl-3-hexene, however, accepting the conventional assignment, he found $R_{polar}/R_{non\text{-}polar} = 0{\cdot}92$ (that is, $\simeq 1/1{\cdot}08$). This, in addition to the fact that the conventional assignment has $T_b(cis) > T_b(trans)$, led Hively to suggest that the conventional assignment is wrong. The ease and rapidity of structural assignment along these lines makes it clear that the method has considerable possibilities.

Up to now, only relatively simple structural problems have been solved chromatographically. Nevertheless, with further development simple generalized rules may be formulated that will greatly reduce the effort now expended in structural assignment.

Studies of diffusion

The process of diffusion occupies a central position in kinetic molecular theory, and explicit equations relating the rate of mass transport (dN/dt) across unit area per unit concentration gradient (dn/dx), to the physical properties of the system may be derived. The precise equations obtained depend upon the model adopted, but all simple approaches start from Fick's laws, the first of which states that

$$dN/dt = -D(dn/dx) \qquad \ldots.(9)$$

The diffusion coefficient D is regarded as being independent of concentration, but being otherwise a function of a number of physical parameters. If a very thin band of material, of amount m, is allowed to diffuse away from the source for time t, Fick's laws yield the expression

$$n/n_{max} = e^{-x^2/4Dt} \qquad \ldots.(10)$$

where $n_{max} = m/2(\pi Dt)^{\frac{1}{2}}$. The distribution of material around the source is thus symmetrical and Gaussian at any time. The obvious similarity between the diffusion bands and chromatographic elution peaks gives a clue to the nature of chromatographic band spreading mechanisms. Equation 10 leads in turn to the very simple and well-known equation

$$\sigma^2 = 2Dt \qquad \ldots.(11)$$

σ being the standard deviation, that is the band half-width at 0·607 of the height. The formal analogy with chromatography is, in fact, developed further because the quantity which characterizes the efficiency of a column, the theoretical plate height H, is defined as σ^2/L, where L is the length of the column.

The direct measurement of diffusion coefficients would seem from the above to be relatively straightforward, but conventional methods are in fact fraught with experimental difficulties. In consequence, the literature contains few data, which is undesirable on theoretical grounds but even more so on experimental ones. A quantitative knowledge of diffusion rates is frequently of paramount importance not only in chemistry but in such fields as physiology, biology, and engineering. Thus, any technique which offers a simple and rapid approach to the measurement of D, even if not capable of the very highest accuracy, is important. Preliminary studies confirm the view that chromatographic techniques fall into this category.

In the simplest instance, chromatographic study of diffusion in the gas phase requires a column of no sorptive capacity for which gas diffusion is the only significant band-broadening process. These conditions are realized in the use of a long (say >3 m) empty stainless steel or glass capillary tube (diameter <1 mm), at gas flow rates below about 1 cm per sec. In an elution experiment the 'chromatographic' peak can be used to evaluate D via equation 11, which for these purposes becomes

$$D=\frac{\sigma^2 u}{2L} \qquad \ldots(12)$$

where u is the carrier gas velocity. Studies of liquid phase diffusion by this method are difficult because of the low rate of diffusion. However, this problem can be overcome by a modification of the method in which, once the diffusing substance has been introduced into the column, the carrier flow is stopped for any desired length of time. Flow is then resumed and the band delivered to the detector. This approach can also, of course, be used with good effect in gas phase studies. If for any reason velocities higher than that suggested are necessary, small corrections for the effect of the parabolic velocity profile in laminar flow [25] and of finite pressure gradient [26] may be necessary. To date, only a few studies of this kind have actually been carried out [26, 27], but the accuracy with which D may be measured can be gauged from Table III.

Table III

Chromatographically measured gas phase diffusion coefficients (D_g) for gases in nitrogen at 1 atm. and 51°C [27]

Gas diffusing	D_g(cm²/sec)		
	Open tube	Packed column	Literature
H_2	0·892	0·886	0·909
O_2	0·263	—	0·244
CO_2	0·188	0·176	0·194

The agreement between the results with open tubes and the values quoted in the literature is as good as can be expected, and it seems likely that, with refinement of procedure and apparatus, very accurate data can be obtained.

As seen in the table, diffusion data may also be obtained with packed columns. In this case, however, the presence of packing contributes to σ, and a correction must be made. The correction methods are established, but become very laborious to apply if any sorption occurs. Even so, in the course of everyday analysis, reasonable estimates of D may be obtained if runs are carried out at several values of u. Useful, if approximate, information of this kind already lies buried in the analytical files of many laboratories.

An alternative approach to the measurement of diffusion coefficients in solution may be based on the detailed study of band shape when sorption occurs. Theoretical equations relating the contribution of liquid phase kinetic processes to σ have been developed (see, for example, [28]). Unfortunately, all the equations contain a quantity, the average liquid film thickness, that is difficult to measure. Reasonable estimates may be made, however, and these usually yield acceptable values of D. The problem is not insurmountable, and is currently under study. However, it may well be that ultimately this approach will be reversed and the method used for measuring liquid film thickness.

Conclusions

This review of the non-analytical uses of gas chromatography has been limited to a few of what now appear to be the most profitable areas of study. There are, however, others that may well develop rapidly in the near future. For example, studies of complexing equilibria have been carried out, and, in at least one case, with remarkable success [29]. Again, columns which act both as a catalyst bed and as an analyser have shown promise in the study of heterogeneous reactions. In a different direction, the possibility of chromatographic measurement of high-polymer molecular weights has been suggested, and specialized techniques for surface-area measurement have been developed. Whether or not these studies individually come to fruition, it seems certain that chromatographic methods will eventually take their place among the conventional techniques of physical chemistry. It would not be surprising if this field, at present undeveloped, expands as explosively as did the analytical side of gas chromatography during the last decade.

References

[1] James, A. T. and Martin, A. J. P. *Biochem. J.*, 50, 679, 1952

[2] van Deemter, J. J., Zuiderweg, F. J., and Klinkenberg, A. *Chem. Engng. Sci.*, 5, 271, 1956.

[3] Glueckauf, E. *Trans. Faraday Soc.*, 51, 34, 1955.

[4] Hoare, M. R. and Purnell, J. H. *Ibid.*, 52, 222, 1956.

[5] James, D. H. and Phillips, C. S. G. *J. chem. Soc.*, 1600, 1953; 3446, 1954.

[6] Bohemen, J., Langer, S. H., Perrett, R. H., and Purnell, J. H. *Ibid.*, 2444, 1960.

[7] Perrett, R. H. and Purnell, J. H. *J. Chromatog.*, 7, 455, 1962.

[8] Reilley, C. N., Hildebrand, G. P., and Ashley, J. W. *Analyt. Chem.*, 34, 1198, 1962.

[9] Ashworth, A. J. and Everett, D. H. *Trans. Faraday Soc.*, 56, 1609, 1960.

[10] Everett, D. H. and Stoddart, C. T. H. *Ibid.*, 57, 746, 1961.

[11] Hofstee, T., Kwantes, A., and Rijnders, G. W. A. *Proc. Int. Syrup Dist.*, p. 105, Brighton. 1960.

[12] Freeguard, G. F. and Stock, R. 'Gas Chromatography' (edited by R. P. W. Scott), p. 102. Butterworths, London. 1962.

[13] Anderson, J. R. and Napier, K. H. *Aust. J. Chem.*, 10, 250, 1957.

[14] Martin, R. L. *Analyt. Chem.*, 33, 347, 1961; 35, 116, 1963.

[15] Burnett, M. R. *Ibid.*, 35, 1567, 1963.

[16] Evered, S. and Pollard, F. H. *J. Chromatog.*, 4, 451, 1960.

[17] Desty, D. H. and Swanton, W. T. *J. phys. Chem.*, 65, 766, 1961.

[18] Martire, D. E. *Analyt. Chem.*, 33, 1143, 1961; 'Gas Chromatography' (edited by L. Fowler), p. 33. Academic Press. New York. 1963.

[19] Langer, S. H. and Purnell, J. H. *J. phys. Chem.*, 67, 263, 1963.

[20] Bell, R. P. *Trans. Faraday Soc.*, 33, 496, 1937.

[21] Ogilvie, J. D., Simmons, M. C., and Hinds, G. P. *Analyt. Chem.*, 30, 25, 1958.

[22] Beaven, G. H. 'Vapour Phase Chromatography' (edited by D. H. Desty), p. 31. Butterworths, London. 1957.

[23] Beaven, G. H., James, A. T., and Johnson, E. A. *Nature, Lond.*, 179, 490, 1957.

[24] Hively, R. A. *Analyt. Chem.*, 35, 1921, 1963.

[25] Golay, M. 'Gas Chromatography' (edited by D. H. Desty), p. 35. Butterworths, London. 1958.

[26] Giddings, J. C., Seager, S. L., Stucki, S. R., and Stewart, G. H. *Analyt. Chem.*, 32, 867, 1960.

[27] Bohemen, J. and Purnell, J. H. *J. chem. Soc.*, 360, 1961.

[28] Purnell, H. 'Gas Chromatography', pp. 125, 142. Wiley, New York. 1962.

[29] Du Plessis, L. A. and Spong, A. H. *J. chem. Soc.*, 2027, 1959.

[30] Cornforth, J. W., Cornforth, R. H., and Mathew K. K. *Ibid.*, 112, 1959.

The learning of discriminations by animals

N. S. Sutherland

Previous experiments on learning in animals have produced some paradoxical results that have never been satisfactorily explained. This article describes a new model of the learning process that attempts to infer the logic of the system mediating learning without specifying how the elements composing the system work from a physiological standpoint. Such a model can be tested only by computer simulation, and the results of the first steps towards this are described.

This article describes an attempt to construct a new model of discrimination learning. Since psychologists use the term 'learning' in a technical sense, it may be as well to start by attempting a rough definition. Learning is the process mediating changes in an animal's responses, where these changes occur as the result of previous experience: by experience, we mean not mental events but any sequence of stimuli and responses. We say that an animal discriminates between two stimuli when it consistently gives one response to one stimulus and a different response to the other. For the present purposes, inactivity is counted as a response, so that if an animal stays still in the presence of one stimulus and moves forward in the presence of another, it would be said to be giving different responses to the two stimuli and therefore to be discriminating between them. 'Discrimination learning' is the process whereby an animal learns to give consistently different responses to two or more stimuli. Most, if not all, learning involves discrimination. It may help to clarify what is meant by 'discrimination learning' if we describe a piece of apparatus used to train rats to make visual discriminations.

Figure 1 illustrates a 'jumping stand': the apparatus was invented by K. S. Lashley in 1930. A rat is placed on the jumping platform (J) and has to learn to jump to one or other of the ledges (L) underneath the two stimuli, which in the apparatus illustrated are horizontal and vertical stripes. These stimulus patterns are attached to doors, and the rat must learn to jump to the ledge under the door bearing the 'positive' stimulus. Which stimulus is positive is determined by the experimenter. The doors can be interchanged and the side on which the positive stimulus appears is changed in a random way from one trial to the next. If the rat jumps to the positive stimulus, it finds the door bearing it unlocked and can push it down and obtain food from the space behind (F). If it jumps to the negative stimulus, it is left for 5 sec on the ledge in front of it, and then removed from the apparatus: a partition (P) prevents the rat moving across from one ledge to the other. The tests are run when the rats are hungry, and the food they obtain is said to be a 'reinforcement'; in general, a reinforcement is any operation that increases the probability of recurrence of the preceding response. Discrimination learning is involved, since the rat learns to make one response (jump towards) to the positive stimulus, another (avoid) to the negative stimulus. One jump constitutes a trial. This apparatus produces very successful learning of visual discriminations by rats. Three features are worth noting. Because the animal has to jump, it hesitates before it makes a response, and this gives it time to see the stimuli between which it has to discriminate. Because jumping is visually controlled and the rat looks towards the spot to which it jumps, the stimuli are seen on each trial in central vision. There is a minimal delay between making the response to the stimuli and obtaining the reinforcement (food in the food compartment).

Three well-known but puzzling phenomena will now be described in order to illustrate what facts a model of discrimination learning must explain.

1. *The overtraining reversal effect.* If rats are trained to jump to a horizontal rectangle and not to a vertical rectangle, the more training they are given at this stage, the faster they subsequently learn to jump to the vertical rectangle and not to the horizontal. The second stage is known as 'reversal learning', since the animals have to learn to reverse the responses learned at the first stage. The results have an air of paradox, since one might expect, and early theories of learning predicted, that the more reinforcements a rat received for one choice, the more difficult it would be for the animal to reverse its responses.
2. *Transfer along a continuum.* Suppose we want to train rats to discriminate between doors coloured in two neighbouring shades of grey. The most efficient method is first to train them

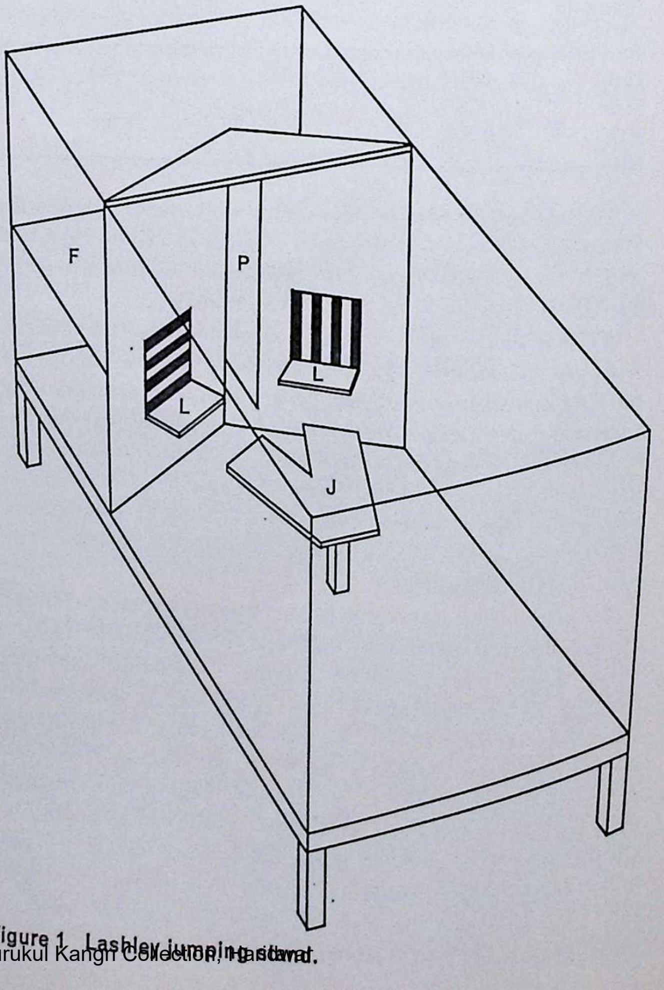


Figure 1 Lashley jumping stand.

N. S. Sutherland, *M.A., D.Phil.*

Was born in 1927 and educated, first as a classicist and then as an experimental psychologist, at Oxford. He has been doing research on perception and learning there since 1953, apart from one year spent at the Massachusetts Institute of Technology. He is at present lecturer in experimental psychology at Oxford University and Fellow of Merton College, and Professor-elect at the University of Sussex.

to discriminate between black and white doors, and then to train them with doors that approximate more and more closely to the two neighbouring shades of grey, keeping either the brighter or the darker door as the positive stimulus the whole time. If we train them with the two neighbouring shades of grey from the outset, it takes them many more total trials to master the discrimination. Again there is a paradox involved: the quickest way of training rats to discriminate between neighbouring shades of grey is not to reinforce them all the time on the stimuli we want them to discriminate, but to reinforce them first on other stimuli that have a larger difference along the same dimension—brightness in this case.

3. *The partial reinforcement effect.* If a rat is being trained to run down an alley for a piece of food at the end, we can either leave food there on every trial or leave food there on half the trials only. The first procedure is known as continuous reinforcement, the second as partial reinforcement. If we now train two groups for 100 trials with the two procedures, and then run them without leaving food at the end of the alley, both groups will eventually cease to run: this procedure is known as extinction. The interesting thing is that the partially reinforced animals continue to run for many more trials of extinction than the continuously reinforced animals, although in their original training they received only half the number of reinforcements. Once again there seems to be a paradox. One might expect that the more a habit was reinforced, the stronger it would be, and the longer it would take to extinguish when reinforcement was not given.

Although the above three results have been described in terms of highly specific situations and stimuli, they have very considerable generality: they have been obtained with many different species, using different training situations and different stimuli. All three results are well known, but none of them has been adequately explained. There are two main reasons for this failure.

The phenomena of animal learning have been studied mainly by American behaviourists, such as C. L. Hull, and they concentrated on the problem of how new responses are learned. However, it will be argued here that perceptual processes play a vital part in learning and that the phenomena of learning cannot be fully understood without taking perceptual processes into account. Secondly, the behaviourists were so anxious to avoid vague explanations in terms of mental processes that they tried to postulate the minimum number of processes intervening between the stimulus and the response. In particular, they tried to avoid the use of the concept of 'attention' which they took to be mentalistic, and held that when a response was learned it was attached to all the stimuli impinging on an organism at the time at which the learning took place.

In recent years a new type of theorizing has become current in psychology. Workers have tried to infer from behavioural data the logic of the system mediating the behaviour. This type of approach has been stimulated partly by the development of computers and other information-processing machines. In order to understand the logical design of a computer, one does not necessarily have to understand how all the basic elements work—one can use relays, electronic valves, transistors, a system of levers, or a hydraulic system to build a computer with a given logic. It may, therefore, be possible to understand something about the logic of the brain without necessarily knowing how all the components work, or without even being able to identify the components specified in the logical system. The chances of doing this have been greatly improved by advances in computer technology, since one way of being sure that one has a logic that really works is to simulate it on a computer and demonstrate that the input–output relationships obtained correspond to those observed in animals.

It is sometimes argued that it is impossible to discover the logic of the brain except by studying its structure and the functioning of its elements, and that for any known relationships between input and output, there could be an infinity of logical systems. While it is true that the study of structure and of the functioning of individual elements may yield very important clues to the logic of the system, there are unfortunately many problems in experimental psychology where these studies give very little help. Moreover, it is often forgotten that although there may be many alternative logics producing the same input–output relationships, the problem at the moment is to think of even one really plausible one that can be tested by predictions from it about further behavioural phenomena. This article describes an attempt to construct such a model.

Description of the model

The model was suggested by the author, and N. J. Mackintosh has played a large part in developing and testing it. Some of the basic ideas derive from the work of K. S. Lashley and I. Krechevsky in the thirties, and more recently D. H. Lawrence performed an ingenious series of experiments with similar underlying ideas [3].

The model can best be understood with reference to figure 2, which is to be read from the bottom upwards. We assume that the stimulus input goes into a number of analysers (represented by circles). The outputs given by each analyser vary with variation of the stimulus along a single dimension only. For example, in the case of vision, one analyser might detect differences in brightness; and differences in its outputs would be determined solely by the shade of black, white, or grey of the stimulus. A separate analyser might detect differences in the ratio of horizontal to vertical contour in a visually presented shape; a third might detect amount of contour; and so on. The ability of all higher animals to 'abstract' features of the environment implies that at some level of the nervous system some such arrangement must exist: octopuses, goldfish, chickens, rats, monkeys, and human beings can all learn to give one response to a horizontal rectangle and another to a vertical rectangle, independently of the size or brightness of the rectangles.

Any output from any analyser can become attached to any response as a result of learning. In figure 2, the left-hand analyser detects brightness differences in the stimulus. If the

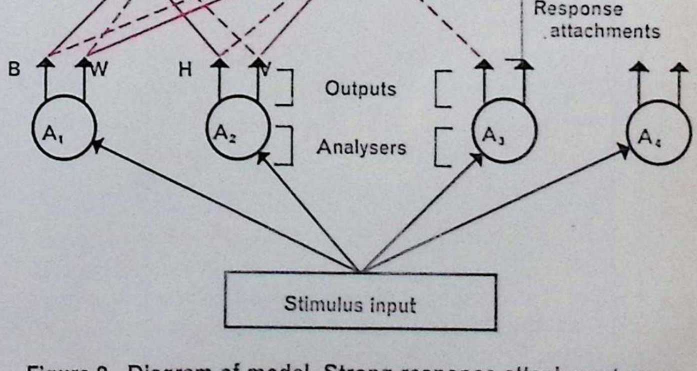


Figure 2 Diagram of model. Strong response attachments are indicated by solid red lines; some possible further response attachments are indicated by dashed red lines.

stimulus is black, output B is given; if it is white, output W is given. If output B occurs, response R_1 is made, and response R_2 is made to output W. Similarly, analyser 2 detects differences in the amount of horizontal or vertical contour present in a stimulus (outputs H and V). The significant feature of this model is that not all analysers are used simultaneously. Only when an analyser is switched in will its output determine which response occurs. Thus in figure 2, if the input is a vertical black rectangle, response R_1 will occur if analyser A_1 is switched in, but response R_2 will occur if analyser A_2 is switched in. Whether or not an analyser is switched in is not an all-or-none process, as different analysers can be more or less strongly switched in. On this model, discrimination learning involves two separate processes: learning to switch in the analyser whose outputs discriminate between the stimuli to be discriminated, and learning which responses to attach to those outputs. The following rules govern the operation of the model.

Rule 1. Changes in the response attachments to the outputs of a given analyser occur only when that analyser is switched in: the more strongly it is switched in, the faster such changes occur.

Rule 2. The more strongly one analyser is switched in, the less strongly are others switched in.

Rule 3. The link between an output and a response will be strengthened whenever an animal is reinforced after the occurrence of that response to that output.

Rule 4. The strength with which an analyser is switched in increases only when different outputs are consistently followed by different events of importance to the animal. It follows that an analyser will not be strengthened if reinforcement is given irrespective of what outputs that analyser gives: it will be strengthened maximally if an animal is always reinforced after one output occurs and never reinforced after a different output.

Rule 5. The strength of a response and also the strength with which an analyser is switched in reach a maximum. The strengths of response attachments reach this maximum more quickly than does the strength with which an analyser is switched in.

It should be noted that changes in the strengths of response attachments are governed by rules different from those governing changes in the strengths of analysers (Rules 3 and 4). This difference yields some interesting predictions which will be discussed below, but the difference is plausible. To survive, animals in nature must do things that lead to reinforcement (Rule 3). They must also analyse reliably those aspects of the stimulus situation that predict different subsequent events. Thus an animal needs to pick up stimuli that signal danger, even though the reception of such signals is not followed by reinforcement. Moreover, Rule 4 prevents an animal storing redundant information and thus reduces the storage load. There is considerable evidence that redundant information is not stored.

The above statement of the model is sketchy, and some assumptions have been omitted, particularly those dealing with the ways in which response attachments and analysers are weakened by non-reinforcement. How the model works should become clearer in the next section, where we shall explain the three puzzling phenomena described above on the basis of the model, and also give some predictions derived from it.

Some explanations and predictions

The explanation of the overtraining reversal effect can most easily be understood by referring to figure 3. The left-hand graphs are schematic learning curves for animals that receive few training trials (Group F), and the right-hand curves are for animals receiving many training trials (Group M). The bottom graphs represent changes in the strength of the relevant analyser, and the middle curves are for changes in the strength of the response attachments to the outputs of that analyser. R_1 represents the response of jumping to the stimulus which is positive at the first stage, and R_2 is the response of jumping to the stimulus which is positive during reversal. The top curves are the proportion of trials on which R_1 and R_2 are given at different stages of training and reversal, and these curves are a joint function of the lower curves. By rule 5, the additional training trials given to Group M will strengthen the relevant analyser more than the correct response. We now consider what

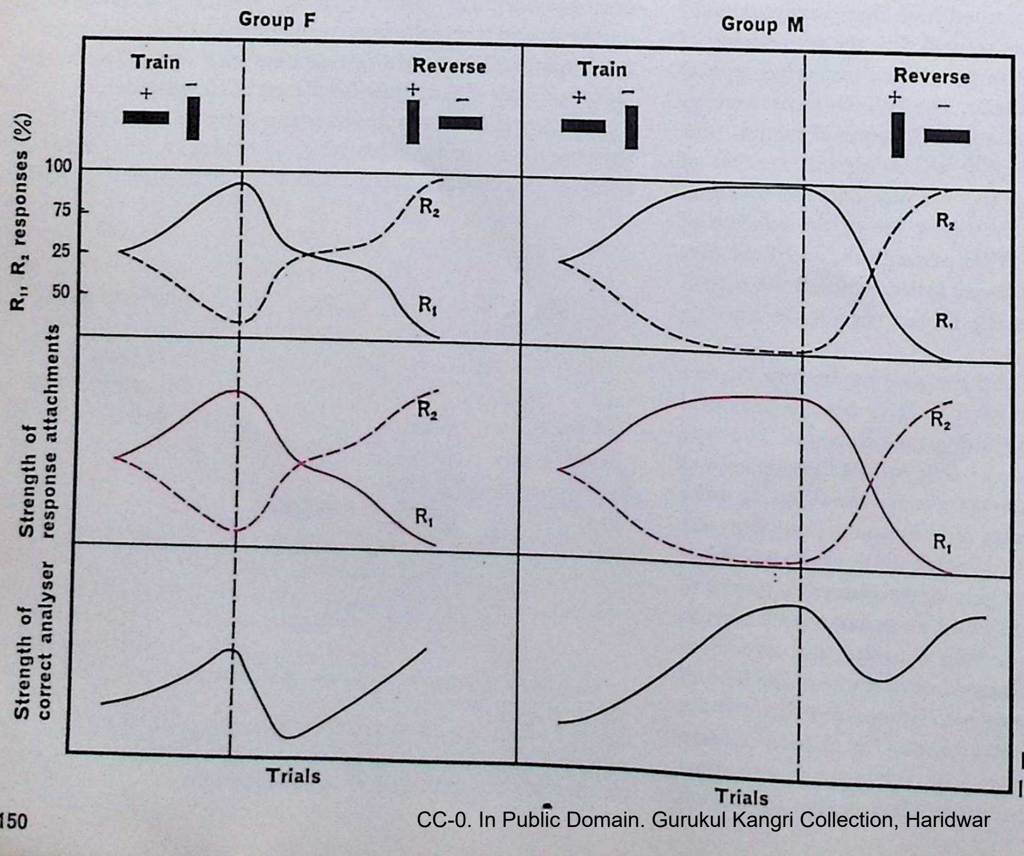


Figure 3 Theoretical reversal learning curves.

will happen in reversal, that is in the trials shown to the right of the dotted vertical line in figure 3. When animals cease to be rewarded for jumping to the horizontal shape, the strength of the correct analyser and of its response attachments will weaken. Since in Group F the analyser is not very strongly switched in, it will rapidly weaken, and when it does the learning of the new responses will be slowed up (by rule 1). Only as the new responses are learned will the analyser again increase in strength, and this means that there will be a period during which animals respond at about chance level while their responses are mainly controlled by other, irrelevant, analysers. In Group M, on the other hand, the analyser is in strongly at the end of training, and the original response attachments to its outputs will therefore be decreased more quickly in comparison with Group F. More incorrect responses will be made at the beginning of reversal, as the analyser is more strongly switched in. The analyser strength will also be decreased, but provided it is still switched in sufficiently when the original responses have been detached, the new response learning can occur at once. As this happens, the analysers will again be strengthened and the total number of trials to learn the reversal will be less than for Group F. That this is the typical course of reversal learning for animals receiving different amounts of training at the first stage is shown by the black curves in figure 4. These are taken from an experiment performed on rats by Mackintosh.

If this account is correct, it follows that if animals are first trained on a discrimination, such as that between black and white, that requires a different analyser from that to be employed at the second stage, for example between horizontal and vertical, then the more training given at the first stage, the slower the animal will be to learn the second stage. The extra training at the first stage now strengthens an analyser not required at the second, and thus makes it more difficult for the animal to switch in the correct second-stage analyser. This prediction was tested and confirmed by Mackintosh [1].

A similar explanation can be given of transfer along a continuum (2 above). If animals are trained from the outset on a difficult discrimination, it will take many trials for them to find the right analyser. Because the two greys are very close together, the appropriate analyser cannot resolve them completely—there will be some overlap in the outputs for the positive and negative stimuli, and hence the outputs can never predict reinforcement and non-reinforcement with complete consistency. It will therefore be difficult for the animal to find the correct analyser. If, however, it is first trained with black and white (which involve the same analyser as the two greys), there will be no overlap between the outputs at this stage, and hence the correct analyser will be rapidly strengthened. The most effective way of training animals to discriminate neighbouring greys will therefore be to start by training them to discriminate black from white.

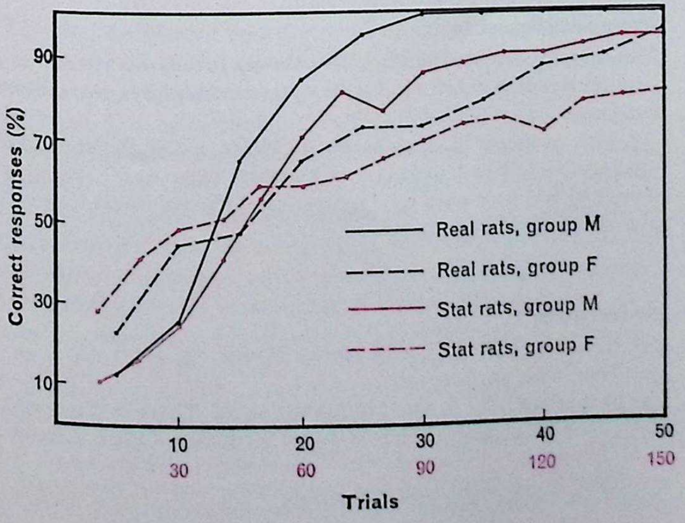


Figure 4 Experimental reversal learning curves.

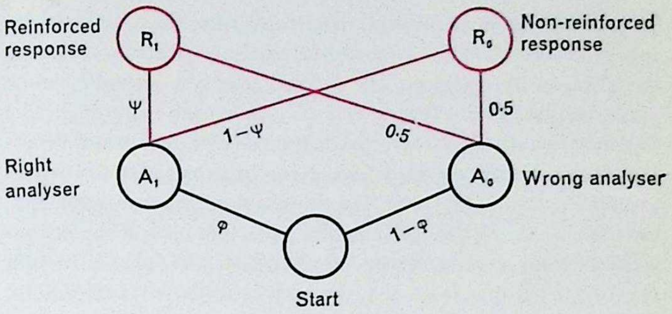


Figure 5 State diagram of Lovejoy's model.

The model also explains the partial reinforcement effect (3 above). If a rat is reinforced for responding on every trial, the response can be attached to the outputs of one analyser or, at most, a small number of analysers: whichever analyser the rat uses initially will predict reinforcement with complete consistency. If however, the response is randomly reinforced on half the trials, no analyser can consistently predict reinforcement: therefore many different analysers will be tried out, and the response should be attached to the outputs of all these analysers. Since the continuously reinforced rats have attached the response to a small number of analysers, they will stop running as soon as the response to these analysers is extinguished or as soon as the analysers themselves are switched off. Partially reinforced rats, on the other hand, have the response attached to many analysers, and must either switch all these analysers off or extinguish the response attachment to each of them in turn: hence they will continue to run for longer. We have undertaken experimental tests of this explanation and have confirmed directly the prediction that partially reinforced rats will attach responses to more cues than continuously reinforced animals.

Thus the model explains the three puzzling findings with which we began. It explains many others as well, but these three were chosen as illustrations because they are such diverse phenomena. There is at first sight no link between them. The diversity of the phenomena a model will explain is of course an important criterion of its usefulness. The model has also led to many new predictions [2] of which we have so far tested about a score. One must suffice here. If we train animals to jump to a white horizontal rectangle and not to a black horizontal rectangle, they can solve the problem by learning to jump to white, or by learning to jump to horizontal, or by learning both. The model predicts (Rule 2) that the more an individual animal learns about the brightness of the rectangles, the less it will learn about their orientation, and this has been found to be true [4].

Formalization of the model

In so far as we have not specified exactly the function relating analyser and response strength to the animal's actual performance in training, the model is rather loose. This defect can be overcome only by formalizing the model.

The first step has been taken independently by E. P. Lovejoy of the University of Pennsylvania. Some of his work, given in a private communication, will be briefly described. Figure 5 is a state diagram of the formal model that he proposed. At the start of each trial, an animal switches in either the correct analyser (goes into state A_1) or a wrong analyser (state A_0): having switched in an analyser, the animal selects a response which again can be either correct (state R_1) or wrong (state R_0). The values given on the lines connecting the different states represent the probability of an animal going from one state to another on the *n*th trial of training. On the simplifying assumption that only one analyser can be switched in on one trial, if the probability of the correct analyser being in is φ, the probability of a wrong one being in is $1-\varphi$. If the correct analyser is switched in, the probabilities of the correct and of the wrong response being made are defined respectively as ψ and $1-\psi$. If an incorrect analyser is in, the probability of both a correct and incorrect response being made is always 0·5, since the outputs from the wrong analyser are randomly correlated with reward.

Lovejoy also defines operators that change the values of φ and ψ. They are changed after each trial, and the changes made depend upon the four possible outcomes of that trial, that is, upon which of the four possible paths in the state diagram has been followed. The direction of the changes on trial $n+1$ depends upon the outcome of trial n; the details of these changes are shown in the table. Thus if states A_1, R_1 occur on the *n*th trial, the value of both φ and ψ increases on trial $n+1$. The operators used ensure that φ and ψ can take only values between 0 and 1. To summarize the table, the probability of the correct analyser being switched in is increased when an animal is reinforced on a trial on which it is switched in, and decreased after a trial when it is switched in and the wrong response occurs: the same rule governs changes in the probability of a correct response being made given that the correct analyser is in. The probability of the correct analyser being switched in is increased after a trial on which the wrong analyser was selected and wrong response made, decreased if the wrong analyser is selected but the right response made.

States on trial *n*		Direction of change of φ and ψ on trial $n+1$	
		$\Delta\varphi$	$\Delta\psi$
A_1	R_1	+	+
A_1	R_0	−	−
A_0	R_1	−	0
A_0	R_0	+	0

We can now try to discover what will happen to the values of φ and ψ over a series of trials, given that we choose certain starting values for each, and choose values for the operators. No analytic solution to the problem has been found, since what happens to a given rat on a given trial is determined by random factors, given the constraints imposed by the values of φ and ψ. However, it is possible to use a computer to run 'imaginary rats' (stat rats—the phrase abbreviates 'statistical rats') for a series of trials, generating the states on each trial by a random process within the constraints imposed by the values of φ and ψ. These values are changed according to the outcome of each trial and the values of the operators selected. Lovejoy has done this using the model to predict the outcome of reversal learning where one group of stat rats received 150 trials of original training before reversal, and the other 50. It was found that with certain values of the operators, the group that received overtraining reversed more quickly. The learning curves of his stat rats during reversal are reproduced in red in figure 4, and it is obvious that they follow roughly the same course as Mackintosh's curves obtained from real rats. Lovejoy's ingenious work is of course only the first step towards formalizing these ideas, and his stochastic model is simpler than the loose model that I have proposed, and very much simpler than the actual processes that control discrimination learning in the rat's brain.

Conclusion

In evaluating the model, it must be remembered that it has been possible to present only a few of the tests for it here. Moreover, this article deals with work on the rat only, but we have evidence that very similar processes are at work in chicks and octopuses, though some of the details are different. One way of testing the theory is in fact to work with several species: where a different experimental result is obtained in two species, the model is modified to account for the difference. Such a modification usually predicts further differences in the behaviour of the two species, and the model can then be tested by performing experiments to test the existence of such further systematic differences in behaviour. We have already done some work of this sort on rats, octopuses, and chicks, and the results give us further confidence that the theory is on the right lines.

In the introduction we said that the behaviourists were anxious not to introduce the concept of attention, as it was regarded as a mentalistic term. It should now be clear that the model proposed is in fact a systematic account of one aspect of what we call attention in every day language, although we have tried to avoid mentalistic concepts in the theory. One test of the success of this attempt is to discover by computer simulation whether such a logical system will really produce the relations between input and output that we have claimed for it. This is a formidable task, and in attempting it we shall certainly find that the model has to be modified, although so many experimental tests have been conducted that it would be surprising if the main outlines were not correct.

The final proof that the brain really has the logical design suggested may come from developments in neuroanatomy and neurophysiology. It is of course possible even now to speculate about the identification of some of the elements in the model: thus the reticular formation and some efferent pathways in sensory systems may be part of the mechanism controlling what we have termed the switching in of analysers [3]. The model we have described has the virtue that it has led to some surprising new findings, and if it is basically correct, it must eventually be of assistance in the interpretation of neurophysiological and neuroanatomical work on the brain.

Acknowledgments

I am greatly indebted to N. J. Mackintosh for allowing me to use unpublished data of his, and for the stimulus of his ideas over the last five years.

I wish to thank M. Clowes, G. A. Miller, and J. Z. Young for reading the MS and making many helpful suggestions. The work described was supported by the American Office of Naval Research and the Nuffield Foundation.

References

[1] Mackintosh, N. J. *J. comp. physiol. Psychol.*, 55, 555, 1962.
[2] *Idem. Ibid.*, 56, 842, 1963.
[3] Sutherland, N. S. In 'Mechanisation of Thought Processes', Vol. II, p. 575, Her Majesty's Stationery Office, London. 1959.
[4] *Idem. Nature, Lond.*, 201, 528, 1964.

Structural plant polysaccharides

R. D. Preston

The nature of plant cell walls is of great practical, as well as scientific, importance, as these walls are the basis of many long-established materials. This article describes the methods of investigating these structures. One result has been the discovery of a group of seaweeds in which the structure of the cell wall differs from that of all other plants. The article also discusses the light that knowledge of the structure throws upon theories of the biosynthesis of plant cell walls.

During the present century, and particularly since the second World War, interest in the cell walls of plants has steadily increased. These structures, still considered technically as lying outside the cell, were long regarded as the province of academic botany, and even in botany itself of anatomists rather than of cytologists, physiologists, and biochemists. Nowadays physicists, chemists, biochemists, and technologists the world over are engaged in detailed analysis of cell walls. The reasons are not far to seek. Plant cell walls provide the whole of the paper upon which our civilisation has come so much to depend; in the xylem of trees they provide structural material with properties that make timber virtually irreplaceable in building construction; and in the form, for example, of cotton, linen, hemp, jute, and sisal they still provide much of our clothes, string, ropes, and sacking. Such large commercial interests are involved that it is hardly surprising that research centres in almost all countries are busily engaged upon the study of plant cell walls. Yet the vastly increased academic interest springs from a very different source.

A cell wall is synthesized by cytoplasm as a sheath completely surrounding the protoplast and delimiting an individual unit, the cell. The plant body can grow and develop only in so far as its individual cells can grow and differentiate, and this they can do only through materials passing through the cell wall and by an increase in wall area. The cell wall is therefore, on these accounts, of vital importance. Firstly, it is an outer sheath closely associated with, even if not inextricably mixed with, the outer layers of the cytoplasm which form the first barrier to entry of materials; secondly, it is a semi-rigid envelope that must either grow, or be capable of deformation, if the cell is to grow. Besides this the wall presents, in the lamellae deposited internally upon it during its differentiation, a record of the chemical species deposited by the cytoplasm. The specific organization of these compounds gives continuous information, if this can be interpreted correctly, concerning the structure of the surface layers of the cytoplasm responsible for its synthesis. It is for these reasons that biochemical and biophysical investigations of cell wall structure both in growing and adult cells have become so important.

The study of cell walls is beset with all the difficulties common to enquiries in which physical scientists invade the field of the biologist. Since each cell within a tissue is the site not only of its own activity but of reactions with other cells, it is almost axiomatic that, other things being equal, it is preferable to work on organisms consisting of, or containing, cells growing in isolation. It is mainly for this reason that work in my laboratory, after a long preliminary exploration of higher plants, has concentrated on seaweeds. The wall structure of some seaweeds resembles that of higher plants very closely, and we have found greater variability among the seaweeds than among higher plants. As a consequence, they present a picture more widely, and therefore more firmly, based than would come from higher plants alone. When single cells growing in isolation are fully understood, it may be possible to give a correct appraisal of cells in tissues of higher plants. The present article deals therefore both with those seaweeds that resemble higher plants from the point of view of wall architecture and with those that do not. Reference to higher plants will be made where appropriate.

The common architecture of the wall

When the walls of green plants are chemically cleaned [1], there usually remains behind a material that is still coherent and that retains the shape of the cell wall. This is called cellulose. It was shown long ago by W. N. Haworth to consist in the main of linear chains of β-D-glucose residues linked together through carbons 1 and 4—a β-1,4 linked glucan (figure 1). The botanist O. L. Sponsler and his co-workers, and the physical scientists K. H. Meyer and H. Mark, using X-ray diffraction analysis, demonstrated that these chains are arranged strictly parallel to each other over parts of their lengths and are there regularly spaced in a crystal lattice based upon the unit cell: $a=8{\cdot}35$ Å; b (fibre axis) $=10{\cdot}3$ Å; $c=7{\cdot}9$ Å; $\beta=84°$. The structure accepted by these workers has since that time been modified [2] (figure 2). Even now, it is not in detail completely correct. The crystallites within which this regularity obtains are long and thin, more than 600 Å long, and about 50 Å wide. The width varies according to the source of the cellulose. In those celluloses that allow the relevant determinations to be made, the planes of 6·1 Å spacing (figure 2) tend to lie parallel to the wall surface; the crystallites—which correspond to the micelles of Nägeli—are somewhat thinner than they are wide [3] and must, on account of the orientation of the 6·1 Å planes, tend to lie with their flat faces parallel to the wall surface. In only one cellulose of those so far investigated (the cellulose of *Spirogyra*, a freshwater alga) is the orientation known to be different from this. In 1948, the further significant observation was made that, in the electron microscope, the celluloses of some plants take the form of long thin threads [4]. This observation has since been generalized to cover all natural

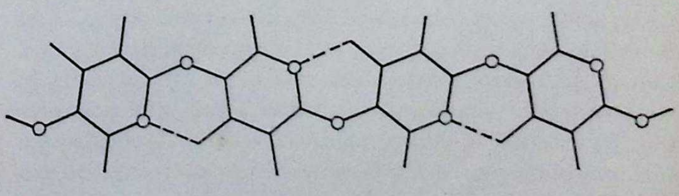


Figure 1 Diagrammatic representation of four β-glucose residues in a cellulose chain. The circles represent the ring oxygens and the linking oxygens, the full lines primary valences, and the broken lines hydrogen bonds between the hydroxyl groups on carbon 3 and the ring oxygen of the next residue.

R. D. Preston, *Ph.D., D.Sc., F.Inst.P., F.L.S., F.I.W.Sc., F.R.S.*
Was born in Leeds, England, in 1908 and was educated at the University of Leeds and at Cornell University, New York. He joined the staff of the Botany Department at Leeds in 1936, and became Professor of Plant Biophysics in 1954. In 1962 he became Head of the Astbury Department of Biophysics at Leeds. His particular research interests are the detailed molecular structure of plant cell walls; problems of water movement in trees; and, more recently, sugar translocation along the stems of trees.

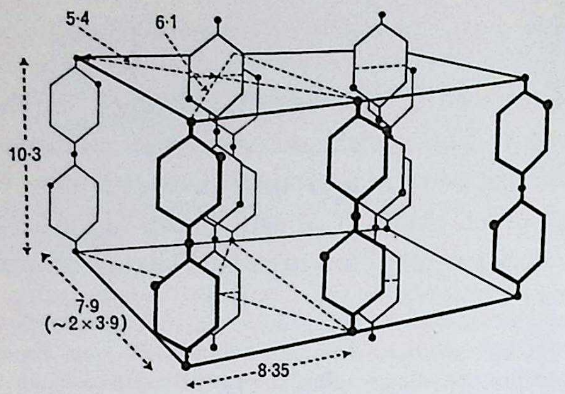


Figure 2 Two unit cells of cellulose, side by side, according to Meyer and Misch. Note that alternate chains lie antiparallel and with a mutual displacement along the chain direction of 10·3 Å. The chains are drawn straight; they should be kinked as in figure 1.

celluloses so far investigated (figure 3). These microfibrils vary in width between about 80 Å and about 300 Å, depending on species, and are approximately half as thick as wide [3]. The crystalline part of the microfibril, responsible for the X-ray diagram, forms a central core within which the parallel and regularly spaced molecular chains are built of β-D-glucose residues only [5]. This core therefore consists of cellulose in the strict sense. The 'cortex' surrounding this core contains chains of sugar residues other than glucose in variable amounts, and this apparently prevents the close packing necessary for crystallinity. The molecular chains in this cortex lie nevertheless presumably parallel to each other and to the chains of the central core, so that this outer layer of the microfibril should be thought of as paracrystalline rather than amorphous. The method whereby cellulose is normally obtained from plants fails to extract these non-glucose sugars, and accordingly the so-called α-cellulose, consisting in the main of microfibrils, rarely hydrolyses to give glucose only. Cellulose from the red alga *Rhodymenia*, for instance, contains about 50 per cent of xylose in its cellulose, and this is by no means an exception [1]. Even α-celluloses of higher plants may contain as much as 15 per cent xylose. The only microfibrils known to consist of glucose residues only are those of the green algae *Valonia*, *Cladophora*, *Chaetomorpha*, and a few other related plants [1, 6].

The microfibrils are embedded in a matrix of substances containing a preponderance of sugars other than glucose which —with occasional exceptions such as the alginic acid [7] of the brown algae—are amorphous or, more probably, paracrystalline. Among these, linear or branched polymers of xylose, mannose, and other sugars are known collectively as hemicelluloses. They have been the subject of much chemical investigation [8]. The other major group of substances in the matrix are the polyuronides; one such is pectic acid, which occurs in the form of partially methylated derivatives and often as the calcium or magnesium salt. The molecular chains of pectic acid are said to be linked both to the hemicelluloses and to cellulose through calcium and magnesium bridges and through phosphate groups. The corresponding substances in the brown algae are the alginic acids. These were originally thought to be poly-D-mannuronic acid, and the X-ray diagram of the crystalline extracted alginic acid was interpreted on this basis. It is now known that alginic acid is a mixture of poly-D-mannuronic acid with poly-L-guluronic acid together with a little polyglucuronic acid [9]. Each of the two main acids can be crystallized separately, and each gives a characteristic diagram. However, the diagram of the mixture extracted from a seaweed is always that of poly-L-guluronic acid, even when the mixture is much richer in the mannuronic component than in the guluronic [12]. Dried brown algae often yield an X-ray diagram of calcium polyguluronate, so that this substance, an incrusting substance, is crystalline within the wall [12]. The diagram proves that the chains of alginic acid lie parallel to the microfibrils. This, together with the demonstration that in higher plants, the xylan chains among the incrusting substances lie parallel to the cellulose microfibrils [10] and other indirect evidence that we need not discuss now, suggests that all linear molecular chains in the incrusting substances of cell walls tend to lie in the same orientation as the microfibrils. They are, then, paracrystalline, not amorphous.

The orientation of cellulose microfibrils

The lamellation of plant cell walls can usually be seen both in the light microscope and in the electron microscope. The finer lamellations often take the form of sheets in which the microfibrils lie on the whole in a closely parallel array. Such lamellae can be seen in the fibres and other elongated cells of higher plants (figure 5) but reach perhaps their highest perfection in some seaweeds (figure 3) [6]. Apart from a few plants (including *Valonia*, in which this structure was first [5*a*] demonstrated), these seaweeds are confined to the order *Cladophorales*, which contains filamentous plants that are sometimes simple as in *Chaetomorpha*, and sometimes branched, as in *Cladophora*. In these seaweeds, the microfibrils of adjacent lamellae are crossed at an angle which is often approximately a right angle. This was demonstrated long before the advent of electron microscopy. Only these two directions occur in *Chaetomorpha melagonium* and *Cladophora rupestris*, so that in these plants the two directions alternate regularly and frequently through the wall. The cells are elongated, and these two sets of microfibrils pass round the cell as helices, a gradual left-hand helix and a steep right-hand helix. In two other species, *Ch. princeps* and *Cladophora prolifera*, there is also a third microfibrillar orientation. In the former the structure is identical with that of *Ch. melagonium* except for the presence of this third set of microfibrils in a moderately steep left-hand helix: in the latter the gradual helix is right-handed, the steep helix left-handed, and this third, moderately steep, helix is also left-handed. Microfibrils are therefore synthesized and oriented by the cytoplasm

Figure 3 Electron micrograph of cellulose microfibrils in lamellae from the side walls of cells of *Chaetomorpha melagonium.* (Magnification 30 000, shadowed Pd-Au.) The microfibrils, some 200 Å wide, lie in two sets crossing at about 90°. Note that one microfibril (centre) turns through 90° from one set to the other.

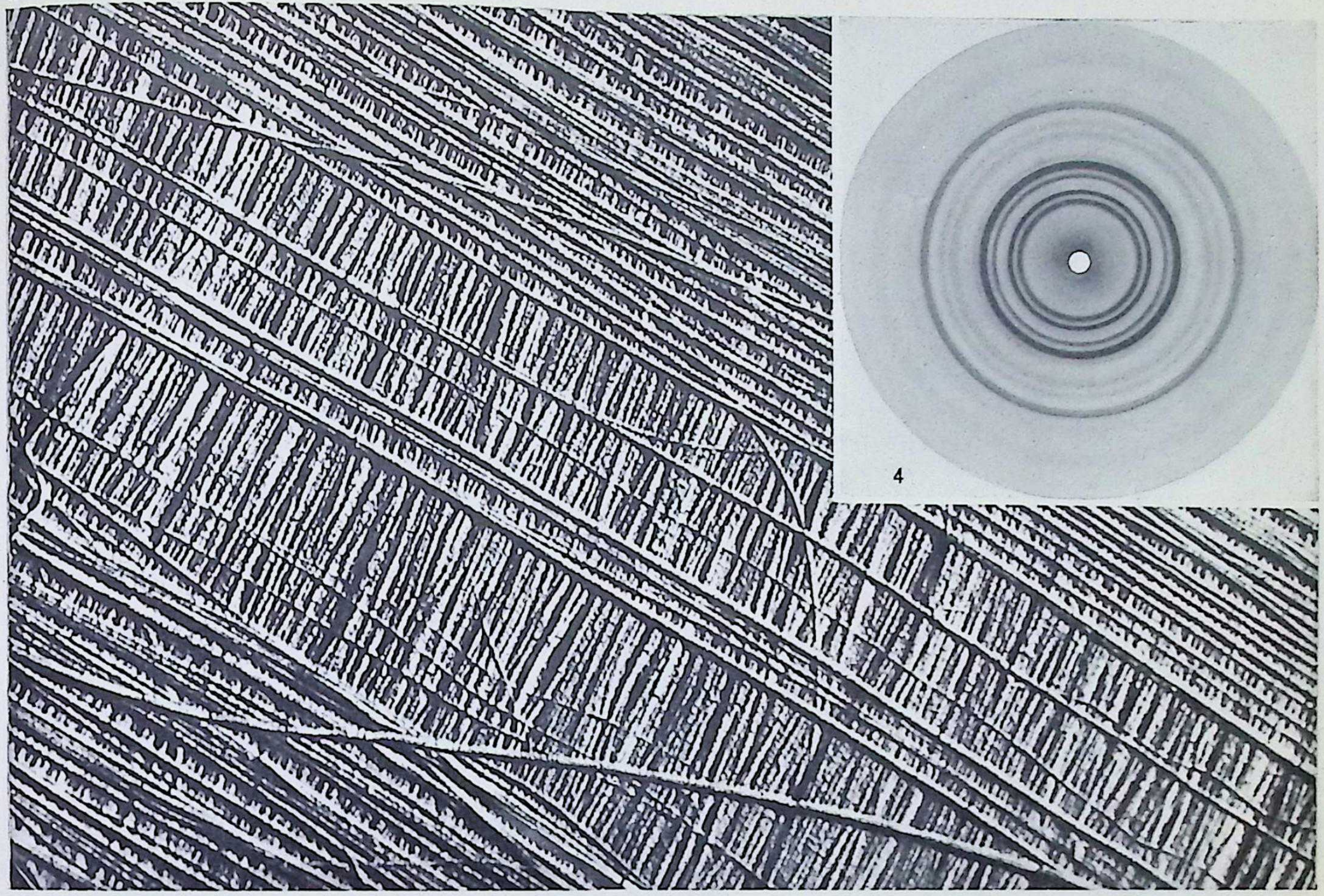
Figure 4 X-ray powder diagram of cellulose (a β-1,4-linked glucan).

Figure 5 Electron micrograph of a carbon replica of the wall of a tracheid from the wood of *Pinus radiata.* Note that in the two lamellae the microfibrils (fine detail) lie in two directions. (Magnification 5100, shadowed Pd-Au.)

Figure 6 Electron micrograph of wall lamellae from *Penicillus dumetosus*, showing xylan microfibrils. Note the microfibrils running vertically (parallel to filament length) with short rod-like bodies lying across them. (Magnification 31 000, shadowed Pd-Au.)

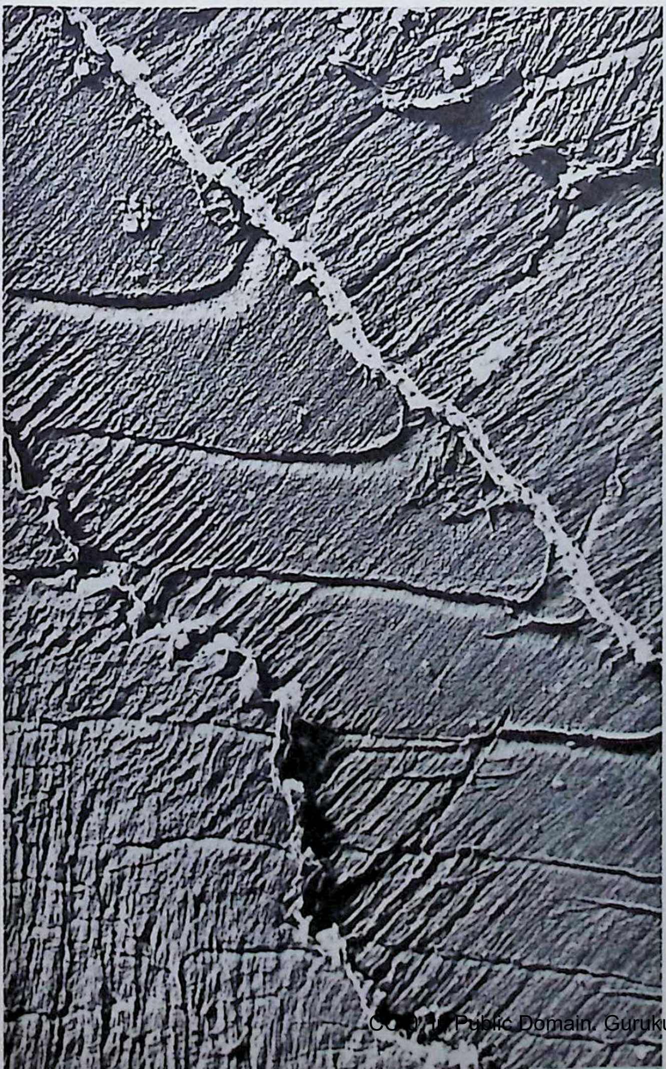
Figure 7 X-ray powder diagram of β-1,3-linked xylan.

3

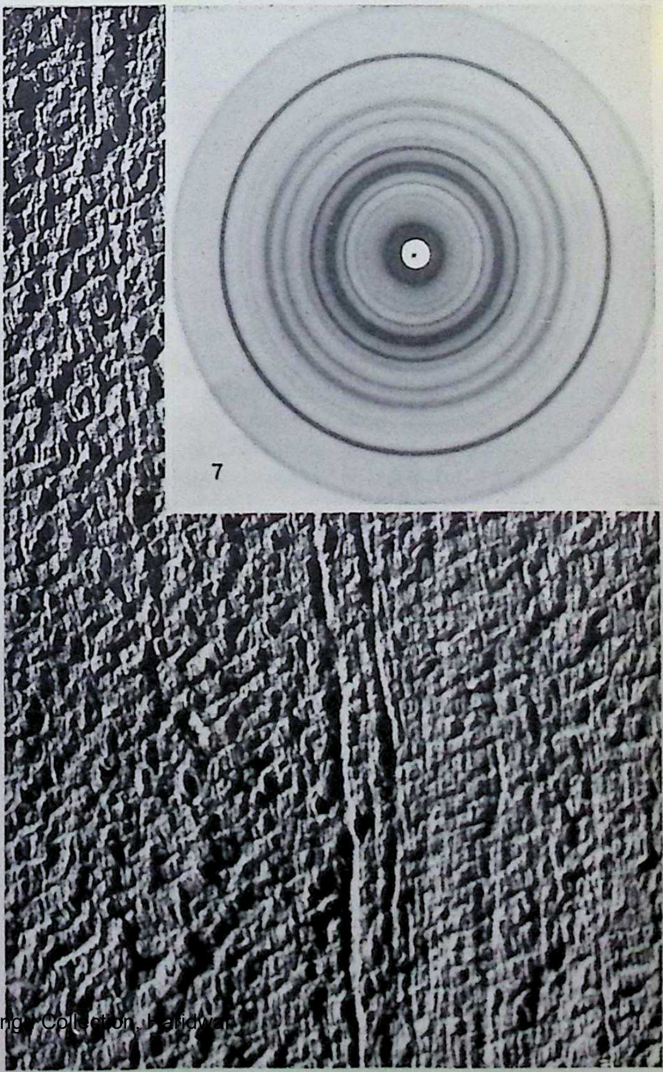


4

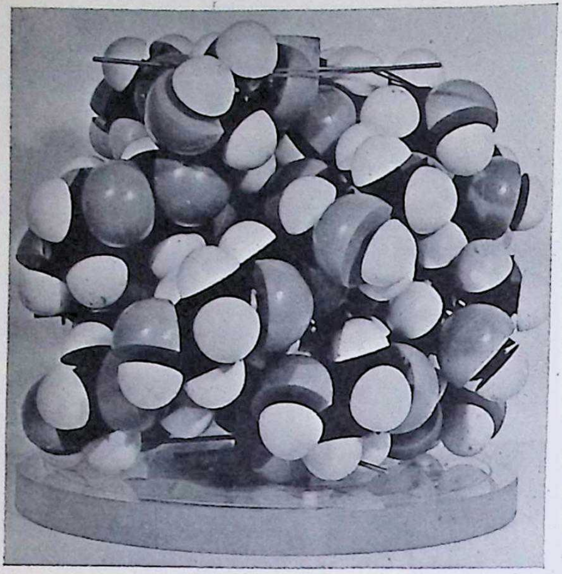
5



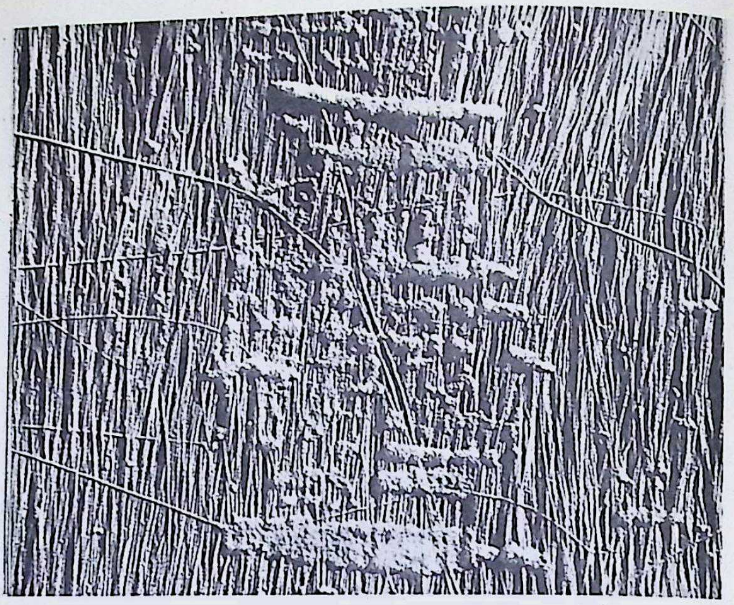
6



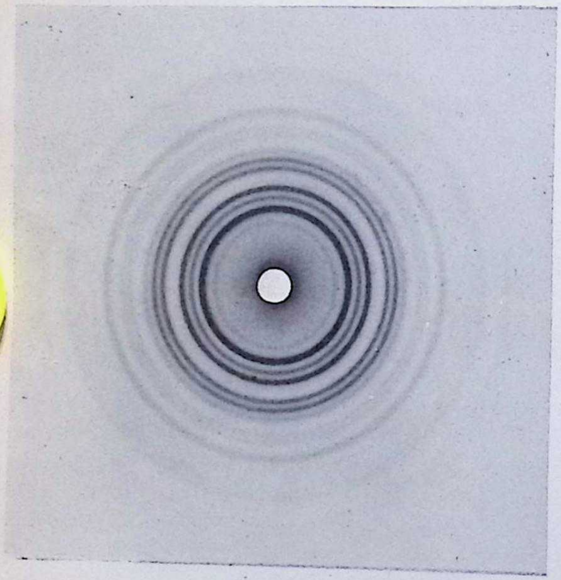
7



8



10



9

11

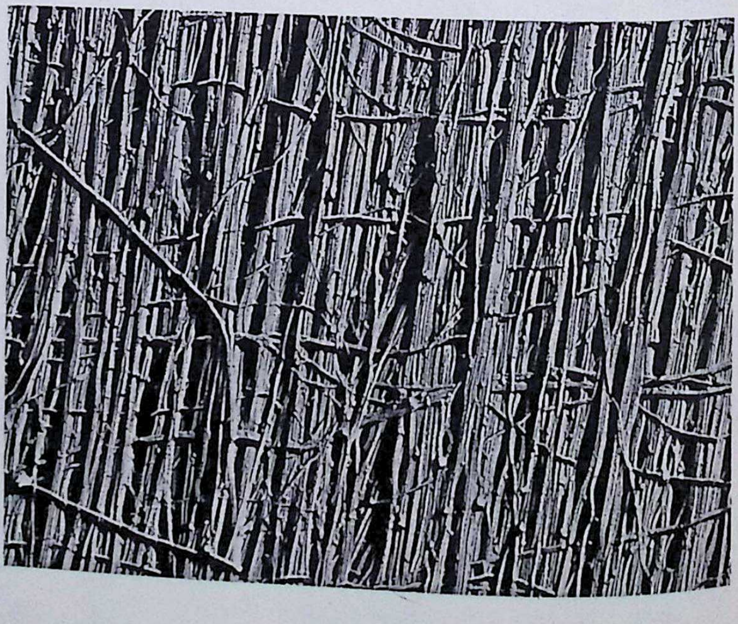


Figure 8 Model of two unit cells of β-1,3-linked xylan, one above the other. The xylan chains are inserted provisionally in two right-handed helices. The white part-spheres represent hydrogen and the grey part-spheres oxygen. One helix begins at centre, bottom, and ends at centre, top. The second helix can be seen running between these points from bottom left to top right.

Figure 9 X-ray powder diagram of mannan (β-1,4-linked).

Figure 10 Electron micrograph of innermost wall lamella from a plasmolysed cell of *Chaetomorpha melagonium.* Note the files of granular aggregates lying almost at 90° to the underlying microfibrils. Microfibrils pass out of these aggregates, apparently representing the first microfibrils of the next wall lamella to be deposited and therefore crossing the microfibrils of the lamella just completed.

Figure 11 Electron micrograph of a wall lamella from a cell of *Batophora* sp. Note complete absence of microfibrils. (Magnification 47 000, shadowed Pd-Au.)

Figure 12 Electron micrograph of lamellae from the cell wall of *Chaetomorpha princeps* (*cf.* figure 3). Note the strong interweaving of microfibrils. (Magnification 20 400, shadowed Pd-Au.)

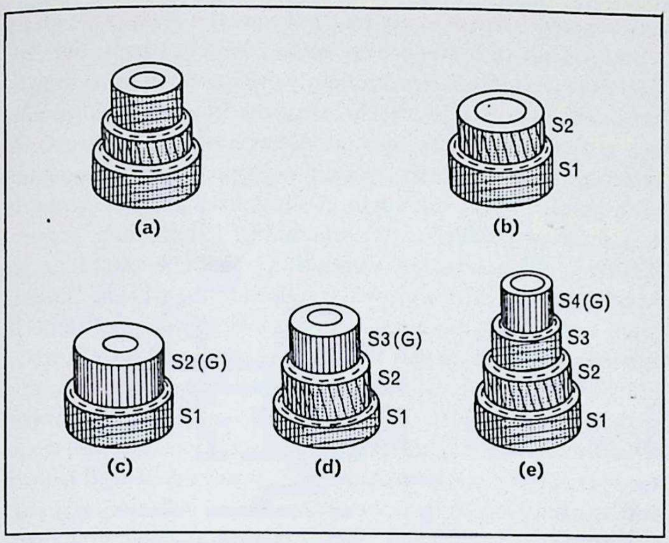


Figure 13 Diagrammatic representation of wall structure in wood tracheids and fibres. The cells are drawn dissected to show S_1, S_2, and S_3 layers. The lines drawn on the lamellae show the general run of the microfibrils. *(a)* Normal three-layered structure of secondary wall. *(b)* Two-layered structure of secondary wall in compression wood tracheid. *(c)*, *(d)*, and *(e)* Two- three- and four-layered structures as observed in tension wood fibres. (By courtesy of Dr A. B. Wardrop, C.S.I.R.O., Australia.)

in these two (or three) directions only, and there is clearly here a most delicate control mechanism at work. The structure resembles in principle the structure of many elongated cells of higher plants (illustrated for the tracheids of wood in figure 13). It seems likely that the derivation of this structure, and the consequences of the structure for the physical properties of the walls, will be closely alike in these seaweeds and in higher plants. One consequence of importance is already known. Each cell, and therefore the whole filament, is compelled to twist as it grows in the phenomenon of 'spiral growth' [6]. This recalls similar phenomena, already described in other plants, which appear to be open to the same explanation [11], though this seems to present an especially clear-cut case. This phenomenon possibly forms the basis of all helical movements in plants, and could explain the 'spiral grain' of conifer wood [12].

Non-cellulosic green plants

It has long been known that in the fungi the microfibrillar component is chitin, not cellulose. Apart from its different chemical nature—which could be regarded as a consequence of the different nutritional circumstances of the fungi—chitin is structurally rather similar to the cellulose wall. Indeed, chitin was at one time referred to as fungal cellulose. It has come to be thought, however, that green plants are always based on cellulosic walls, and current ideas on cell and wall growth, including the action of growth substances, are invariably expressed in terms of a cellulosic skeleton in a matrix of hemicelluloses (mostly xylan and mannan) and pectic compounds. Some oddities have long been recognized. The walls of endosperm cells in palm seeds contain exceptionally large amounts of mannan and very little cellulose; one example is the vegetable ivory used at one time in the manufacture of buttons. The mannan has been said to be microfibrillar [13], but this awaits confirmation; this type of wall could, however, be regarded as an extreme case of the common wall architecture in which the cellulose component is reduced compared with the mannan component. It has, however, repeatedly been claimed to be impossible histologically to demonstrate the presence of cellulose in a number of seaweeds now collected into the orders Caulerpales, Codiales, and Dasycladales. The possibility has therefore for many years been open that here is a group of plants in which the wall organization is different from that commonly accepted.

It has now been demonstrated by modern methods of structure analysis that the walls of these plants are in fact significantly different from those of all other green plants [14]. When the skeletal polysaccharide is extracted from the walls of *Caulerpa*, *Halimeda*, *Udotea*, or *Penicillus*—these are all Caulerpales—it hydrolyses to give xylose instead of glucose. The X-ray diagram (figure 7) is quite different from that of cellulose (figure 4). It is, moreover, not the same as that yielded by the xylan extractable from higher plants. In these plants, therefore, we have a xylan in which the linkage between the units in the chain is different from that found in higher plant xylan. Since the latter is 1 : 4-linked, this new xylan can hardly be other than 1 : 3-linked (figure 14). Over the past three or four years more than one group has shown that all these plants do contain β-1,3-linked xylan in quantity. The X-ray diagram is certainly therefore the diagram of β-1,3-linked xylan. The extracted material giving this diagram is microfibrillar (figure 6), but since the xylan chain is curved (figure 14) the molecular chains within them cannot lie, as they do in cellulose, parallel to the microfibril length. Moreover, the birefringence of these microfibrils is negative, not, as in cellulose, positive. This means that the constituent xylan chains must lie at an angle greater than 45° to the microfibril length.

The fibre X-ray diagram given by parallel arrays of xylan microfibrils can be interpreted, and the birefringence understood, on the basis of double stranded helices (figure 8) in hexagonal close packing. The double helices have been calculated to be 13·7 Å in diameter with a repeat distance of 5·85 Å along the helix axes and with six xylose residues per turn of each helix. From the dry condition, this xylan takes up about 30 per cent of water, and the X-ray diagram is then much sharpened. The helices are then 15·4 Å in diameter and the repeat distance along the microfibril is 6·06 Å.

The internal structures are very different from those of cellulose. The helical form of the constituent chains in xylan make the structure much more open and allows much more extensive hydrogen bonding to water molecules within the crystal lattice, whereas water is unable to penetrate the lattice in cellulose. When this xylan is dried over phosphorus pentoxide, its density makes it necessary to assume that there is water in the lattice to the extent of about one residual molecule per xylose residue. Xylan can be distinguished from cellulose even in the electron microscope, since when cellulose lamellae are torn, there inevitably occurs along the torn edge a fringe of loose microfibrils; these are hardly ever observed on the torn edge of xylan walls. The adhesion between xylan microfibrils appears to be much greater than it is between cellulose microfibrils. The chain coiling, the mutual adhesion between microfibrils, and their high water content suggest that the physical properties of these xylan microfibrils—for instance the mechanical properties—must be markedly different from

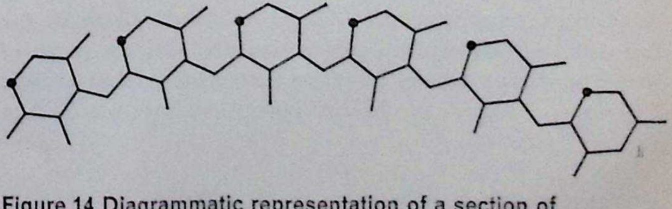


Figure 14 Diagrammatic representation of a section of a β-1,3-linked xylan chain. Note the pronounced curvature.

those of cellulose microfibrils. The relationship between cell growth and wall organization might therefore be different from that obtaining in cellulosic walls. These effects have not, however, so far been sought. The incrusting substances of these walls contain a considerable amount of glucan, at least part of which is 1,3-linked. In the higher plants, a glucan forms the skeletal polysaccharide and xylan is found in the incrusting substances; we have here the reverse condition. In general, however, these walls are architecturally similar to cellulosic walls in being microfibrillar, and the differences might be of degree rather than of kind.

In another group of green plants, largely members of the Codiales, the difference in wall architecture goes even further [14]. If we take *Codium* itself as an example, the products of hydrolysis of the whole wall are mannose, together with a little glucose, and almost nothing else, whereas most green plants, including higher plants, have, combined in their walls, a whole range of sugars and sugar derivatives. The mannose units in the wall are united into linear chains by 1 : 4 links, so that the resulting mannan in this respect closely resembles cellulose. By analogy with other plants, one would expect to find present crystalline and non-crystalline components, and correspondingly the walls do give sharp X-ray diagrams (figure 9) [14], though often only after the wall has been treated in some way, even if only by hot water. As in other plants, the crystallites are sometimes well oriented parallel to each other and then give a sharp fibre diagram. Since the crystalline component is mannan and the incrusting substance is largely mannan, it is not possible to isolate the skeletal substance satisfactorily. It has therefore proved to be impossible to see in the electron microscope the crystalline material under conditions such that it is completely certain that it has not been affected by the extraction treatment. Nevertheless, untreated walls never show any signs of microfibrillar structure as cellulose and xylan walls do. When the walls are treated with, for example, caustic potash solutions, the residue, which is still highly crystalline, appears granular in the electron microscope (figure 12). Walls which give well-defined fibre diagrams might by analogy be expected to contain microfibrils in nicely parallel array but even these have not so far been found to present any but granular appearance. In these plants, therefore, we find walls which are quite exceptional in that they contain no observable microfibrils.

There are a number of seaweeds which contain two crystalline wall polysaccharides, and these may be any two of the trio cellulose, xylan, mannan. *Halicystis*, a green alga resembling *Valonia* in form, contains both cellulose (in the form, however, of cellulose II, mercerized cellulose) and xylan [15], now known to be 1,3-linked [14]. Similarly, *Hydrodictyon*, again a green alga, contains cellulose (this time in the normal form, cellulose I) together with mannan [16]. Perhaps the most intriguing mixture is found in *Porphyra*. This is not only because in the mixture there is β-1,3-linked xylan with β-1,4-linked mannan, but because this is a red alga belonging to a class which is regarded as 'primitive'. In this plant the xylan forms the skeletal polysaccharide of the cell walls and the mannan occurs mainly as an outer sheath which in the whole plant is fused into a 'cuticle' [14]. It is therefore clear that Nature has, as it were, 'experimented' with a variety of sugars in the formation of wall skeletal polysaccharides, and the range of sugars used may indeed be wider than is at present known. Only one of these 'experiments' appears to have survived in higher plants.

The synthesis and orientation of microfibrils

Apart from the members of the Codiales, all plants so far investigated—that is to say by far the greater bulk of all plants examined up to now—possess walls which are microfibrillar. Since the microfibrils are frequently orientated in some specific way, and since the linear chains of the incrusting substances are probably similarly orientated, such walls, anisotropic in structure, must also be anisotropic in physical properties. Walls with parallel microfibrils will, for instance, be more resistant to stretching parallel to the microfibril length than perpendicular to it. If it is true, as seems highly probable, that increase in cell wall area during growth is related to the tension imposed upon the wall by the internal hydrostatic pressure of the cell, it is therefore evident that the shape of a cell will be a function of its wall structure. Dimensional changes during growth, and perhaps even the rate of dimensional change and of growth, should therefore be a function, however complex, of wall structure [11, 17]. It seems probable that processes of wall growth will, qualitatively at least, be similar for all cells similarly constructed. It could well be, therefore, that observations made on, for instance, the seaweed *Chaetomorpha* have significance for the fibrous tissues of higher plants, including wood, in which the cells are too small to allow similar observations to be made directly.

These final paragraphs are based upon the supposition that similar architecture implies closely similar synthetic mechanisms. Although the observations are based largely upon *Chaetomorpha*, the model to be proposed of the synthesizing and orienting mechanism should be of general application.

The relevant observations with the consequent interpretations are as follows:

1. The microfibrils in any one lamella are often twisted around each other. This suggests immediately that the microfibrils are not produced at a plane interface, such as that between cytoplasm and wall, and further that they are not formed by lateral aggregation of preformed chains. They are much more likely to be formed by end synthesis. The microfibrils of the extracellular cellulose of *Acetobacter xylinum* have already been shown to be produced in this way. This type of synthesis in *Chaetomorpha* is further supported by the following observations.
2. Although successive lamellae, with crossed microfibrils (figure 3), can often be separated completely, they are sometimes inextricably entangled by microfibrils of one lamella which weave back and forth between the microfibrils of the next lamella (figure 12) or which turn through 90° to become part of the next lamella (figure 3), or which penetrate the next lamella to be incorporated with a third, remoter lamella. This supports the idea of end synthesis, and also suggests that the microfibrils lying in the two (or three) directions are synthesized simultaneously, not successively. This implies that the synthesizing and orienting mechanism occupies a surface layer of the cytoplasm at least three microfibril diameters thick.
3. After plasmolysis, the inner face of the wall carries granular aggregates lying in files (figure 10) from which pass out microfibrils. The orientation of these files and these microfibrils leaves open the speculation that the granular aggregates may represent parts of a microfibril synthesizing system, which in the living cell would cover the whole cell surface.
4. When the organization of the cytoplasmic surface is disturbed, as it probably is on plasmolysis, and as it certainly is when a cell is preparing to sporulate, then the power of cellulose synthesis is retained but the ability to orient the resulting microfibrils is lost [6]. In both instances the microfibrils produced lie at random.

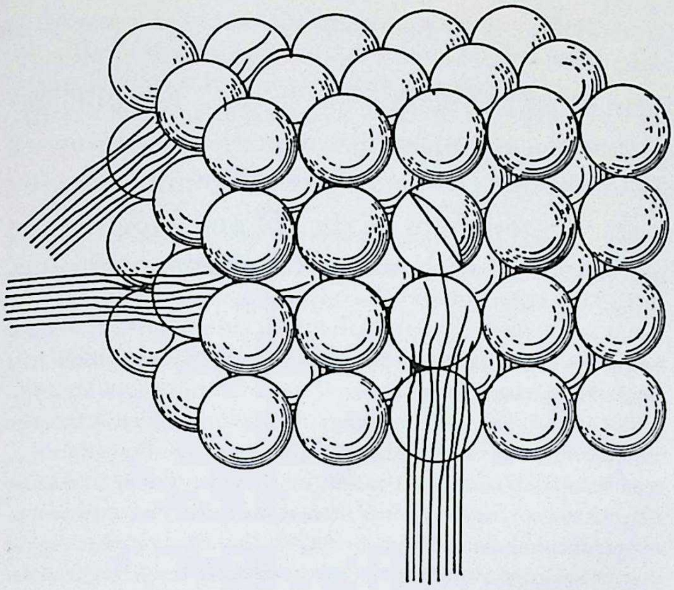


Figure 15 Diagrammatic representation of the proposed model for microfibril synthesis and orientation. The wall surface, lying in the plane of the page, is considered as covered by three layers of granules forming the outer layer of the cytoplasm. Each granule is considered capable of microfibril synthesis when in contact with a microfibril end. Three microfibrils are shown in course of passing through this group of granules. Any microfibril lying in any direction will clearly be constrained, by the cubic close packing of the spheres, to grow in these three directions only. The fourth direction (parallel to the diagonal to the squares, running from lower right to top left) is suppressed by a polarity in the spheres.

From these observations, certain desiderata for a model of the synthesizing system can be laid down:

I. The system must produce mocrofibrils by end synthesis.
II. The system must extend over the whole cell surface.
III. This extensive system must be capable of producing microfibrils in all three directions, either separately or simultaneously, but in no other directions.

The model proposed, in terms of the granular aggregates observed on developing wall lamellae, is illustrated in figure 15 [18]. The cell surface, conceived as lying in the plane of the page, is covered by three layers of granules in cubic close packing, each granule some few hundred Å in diameter. Each granule is considered to represent an enzyme complex which, under the stimulus of a microfibril end as starter, will add glucose residues to the end of a glucan chain. When, therefore, a microfibril end comes into contact with a granule, it is propagated through the granule to the next one, which in turn takes up the synthesis and so forth. Clearly the microfibrils will mostly be constrained to two directions at right angles (as in figure 15), along the sides of the squares, since these directions carry the most dense array of granules. Orientations along the two diagonals would be expected also to occur but to be less common. Only one of these directions is known to occur in the seaweeds and we must assume a polarity in the granules to suppress the other. Since enzyme systems with prosthetic groups are involved, this is not an unlikely assumption.

This model then contains all the essentials appropriate to microfibril synthesis and orientation. It must at the moment be regarded as a model with no real counterpart in the structure of the cell. It has, however, the observational support that closely packed granules have been observed at the surface of *Chaetomorpha* cytoplasm and, more recently and using a most elegant technique, in a similar situation in yeast cells [19]. It does not harmonize with the now classic concept of the plasmalemma delineated in section in the electron microscope as a thin, comparatively smooth membrane. The recent finding, however, that after fixation in glutaraldehyde the place at which this membrane would otherwise have appeared is occupied by a series of tubules [20] seems to be most promising. Whether we find tubules or spheres, there seems here to be the type of organization that wall structure has suggested.

References

[1] Cronshaw, J., Myers, A., and Preston, R. D. *Biochim. biophys. Acta*, 27, 89, 1958.
[2] Meyer, K. H. and Misch, L. *Helv. chim. Acta*, 20, 232, 1937.
[3] Preston, R. D. *Disc. Faraday Soc.*, No. 11, 165, 1951.
[4*a*] Preston, R. D., Nicolai, E., Reed, R., and Millard, A. *Nature, Lond.*, 162, 665, 1948.
[4*b*] Frey-Wyssling, A., Muhlethaler, K., and Wyckoff, R. W. G. *Experientia*, 6, 12, 475, 1948.
[5] Preston, R. D. and Cronshaw, J. *Nature, Lond.*, 181, 248, 1958.
[6] Frei, Eva and Preston, R. D. *Proc. roy. Soc.*, *B*, 154, 70, 1961; *Ibid.*, 155, 55, 1961.
[7] *Idem. Nature, Lond.*, 196, 130, 1962.
[8] Hirst, E. L. *Kolloid. Ges.*, 18, 104, 1958; *Proc. roy. Soc.*, *A*, 252, 287, 1959.
[9] Fischer, F. G. and Dörfel, H. *Z. physiol. Chem.*, 302, 186, 1955.
Drummond, D. W., Hirst, E. L., and Percival, E. *Chemy Ind.*, 1088, 1958; *J. chem. Soc.*, 1208, 1962.
Whistler, R. L. and Kirby, K. *Z. physiol. Chem.*, 314, 46, 1959.
[10] Liang, C. Y., Bassett, K. H., McGinnes, E. A., and Marchessault, R. H. *T.A.P.P.I.*, 43, 1017, 1960.
[11] Middlebrook, Mavis and Preston, R. D. *Biochim. biophys. Acta*, 9, 32; 9, 115, 1952.
Preston, R. D. *Ibid.*, 2, 115, 1948.
[12] Preston, R. D. *Forestry*, 23, 48, 1949.
[13] Meier, H. *Biochim. biophys. Acta*, 28, 229, 1958.
[14] Frei, Eva and Preston, R. D. *Nature, Lond.*, 192, 939, 1961.
[15] Roelofsen, P. A., Dalitz, V. C., and Wijnman, C. F. *Biochim. biophys. Acta*, 11, 344, 1953.
[16] Kreger, D. R. *Proc. Acad. Sci. Amst.*, *C*, 63, 613, 1960
[17] Probine, M. C. and Preston, R. D. *Nature, Lond.*, 182, 1657, 1958; *J. exp. Bot.*, 12, 261, 1961; 13, 111, 1962
[18] Preston, R. D., in Zimmermann, M. H. (Editor), 'Symposium on the Formation of Wood in Forest Trees'. Academic Press, New York. 1964
[19] Frey-Wyssling, A. *Schweiz. landwirt. Forsch.*, 2, 103, 1963.
[20] Ledbetter, M. C. and Porter, H. R. *J. cell. Biol.*, 19, 239, 1963.

The language of machines

Sir Edward Bullard

Most machines do only a few things, and the problem of telling them what to do is not difficult. An electronic computer is an unusual machine in that it can undertake a very wide range of operations. This variety of possible behaviour raises very severe problems of communication between the user and his machine and has led to the development of the art and profession of computer programming, which this article describes, together with some of the languages used.

The typical machine is a device, such as a clock or an automatic lathe, that repeats a predetermined sequence of operations over and over again. The use in ordinary speech of phrases such as 'machine-like' and 'machine made' illustrates this idea of monotonous repetition. A degree of flexibility can be obtained through a human operator, as with a hand lathe, or through a controlling mechanism that can be varied, as in the player piano or the Jacquard loom. The player piano, however, plays only notes that have been specified individually, and the Jacquard loom weaves a predetermined pattern.

In the last twenty years, machines have been developed that escape from this rigidity and whose sequence of operations can depend in an arbitrarily complicated way on the progress of the job they are doing. These are the electronic computers. In a sense they are not different from other machines: they are fully deterministic, that is their behaviour depends exclusively on their structure and on the information supplied to them. They can, however, perform tasks not ordinarily associated with machines. For practical purposes a machine that can be made to translate a Russian text and that can also solve a differential equation is different in kind from a musical box that plays 'Pop goes the weasel' because the required sequence of notes has been built into it.

The machine

An electronic computer is only incidentally, and for convenience, electronic, and has functions wider than computation. Its essential features are that it is digital and programmable. A digital device is one that has elements, each of which can be in one of two or more discrete states. For example, a terminal may have a positive or a negative voltage on it, or a ferromagnetic ring may be magnetized in one direction or the other around its circumference. If, as is usual in computers, each element has only two states, they may conveniently be called 0 and 1, and the information in the machine may be thought of as a string of 0's and 1's, each called a 'bit', an abbreviation of 'binary digit'. The bits are combined in a linear sequence to form machine words, which can be stored, one in each of a large number of storage locations. A machine word can be used to represent, in a suitable code, a number or a linguistic word. A number may be coded as a row of bits in many ways. The simplest way is to regard the right-hand bit as representing 0 if it is 0, or representing 1 if it is 1, the next bit to the left as representing 0 or 2, the next 0 or 4, and so on. For example, 111 would represent $4+2+1$, that is 7, and 101 would represent 5. Other common schemes represent numbers such as 213×10^{17} by coding the 213 and the 17 separately and packing them in a single storage location.

The use of such binary storage requires numbers to be converted from the usual decimal representation when they are read into the machine. It might be thought that it would be simpler to use storage devices with 10 instead of 2 states and to store a decimal digit in each. In fact the greater complexity and cost of the decimal devices more than outweighs the larger number of binary ones needed; also binary arithmetic is simpler and faster than decimal arithmetic. In practice, therefore, the representation of numbers by bits is used in most computers, except those intended for work in which the amount of input or output is very large, when the conversion from decimal to binary and back again may occupy more time than the computation.

The number of bits in a storage location is determined largely, by the requirement that it must contain a number with enough digits to meet practical requirements. This number of bits is increasing; a few years ago it was sometimes as low as 12, but recently it has commonly been 36, and in the latest machines it is 48 or 64. If there are too few bits in a word, accuracy may be lost by rounding-off errors during the computation, and if there are too many bits, storage is wasted; 48 seems a satisfactory compromise. The early machines had a few hundred storage locations, and could therefore store a few hundred numbers. Large modern ones can usually store 32 000 numbers, comprising over a million bits, each of which is a separate physical entity. Each storage location is numbered and, by analogy with the houses in a street, these numbers are called addresses.

Strings of letters, such as linguistic words, are represented in an arbitrary code, usually using 6 bits per letter. For example, on the International Business Machines' computers, C is 010111, A is 010101 and T is 110011. The word CAT is therefore represented by 010111010101110011 and occupies 18 bits, half of a 36-bit storage location.

A storage location can also be used to store purely logical 'yes/no' information. For example, one might store a list of students' names in alternate storage locations and in the intermediate locations put a word containing a 1 as its first bit if the student is a member of the football team, or 0 if he is not, a 1 in the second bit if he is over 6 ft tall, and 1 in the third bit if he has red hair. A red-haired footballer less than 6 ft tall would then be represented by 101, and all of them could be picked out from the list by looking for this combination of bits.

Lastly, a machine word may contain an instruction. An instruction is an element of a program. A program controls the sequence of operations in the job being performed; the machine normally runs through the instructions in sequence, taking a few microseconds to carry out each. A program may contain anything from 20 to 20 000 or more instructions.

To economize in equipment, most operations on stored information are carried out in a special storage location called the accumulator. For example, if two numbers are to be added, one is transferred to the accumulator, the addition is performed

Sir Edward Bullard, *Sc.D., F.R.S.*

Was born in 1907 and was educated at Repton School and Clare College, Cambridge. He began research under Lord Rutherford in 1929, working on the scattering of slow electrons in gases. In 1931 he took up geophysics and has done experimental work on gravity, seismology, and the measurement of heat flow on land and at sea. From 1950 to 1955 he was Director of the National Physical Laboratory, and is now Reader in Geophysics in the University of Cambridge. He has written extensively on the origin of the Earth's magnetic field and on magnetohydrodynamics.

in the accumulator, and the answer transferred back to one of the ordinary storage locations; with this arrangement, only the accumulator need have the circuitry necessary for adding.

It is possible to devise many schemes for representing an instruction as a row of bits in a machine. The simplest is to divide the instruction into two parts, one specifying the operation to be performed and the other the address of the storage location concerned. For example, a pattern of bits at one end of the instruction might mean 'transfer to the accumulator the contents of the storage location whose address is specified by the last 15 bits of this instruction' or 'add the number in the specified location to that in the accumulator'. About a dozen different instructions such as these will allow information to be moved about in the machine and arithmetic to be done.

Normally, when one instruction has been carried out, the machine takes the next one in the program. If this were all, the machine would be on the level of a musical box, as it could perform only those calculations whose nature and order were completely specified beforehand. In all but the simplest programs, the course of the operations must depend on what is found as the work goes along. For example, in calculating the sine of an angle θ from the series

$$\sin\theta = \theta - \theta^3/3! + \theta^5/5! \ldots$$

one would wish to cut off the series at the point where further terms have become small enough to be ignored, and the point at which this happens depends on the value of θ, and might not be known when the program was written. As another example, if the program is looking through a list of names for potential members of a cricket team, it may wish to print the name of a boy if he has played in a previous match and scored over 50 runs. Such needs are met by conditional jumps. A conditional jump is an instruction that determines whether some condition, such as a number being zero, is satisfied and, if it is, causes the machine to jump to a specified address and take the instruction it contains as the next instruction. The versatility of a computer, and the impression of judgment that it gives, depend largely on the use by the programmer of conditional jumps to build up very complicated logical structures.

Machine language

The language understood by the machine consists of strings of bits, that is of 0's and 1's. To write and check such binary instructions is intolerable to a human being, and a means must be provided for linking the programmer and the machine in a way that is more acceptable to the former and yet can be digested by the latter. The simplest of such means are the so-called machine languages. In these, each instruction in the program is written by the programmer in a form that is easy for him to remember and check, and a special assembly program is provided that converts the programmer's program into the form needed by the machine. For example, on the Cambridge University machine EDSAC, the two instructions, 'transfer a number from location 200 to the accumulator' and 'add it to the number in location 300' are

```
10f200
12f300
```

The number 10 means 'move the number in the location specified at the end of this instruction to the accumulator' and 12 means 'add the number in the accumulator to that in the location specified at the end of this instruction'. The 'f' separates the parts specifying the operation and the address. The instructions are punched on paper tape by a typewriter that converts each character into a row of up to five holes spaced across the tape. The assembly program reads the tape and converts each instruction to the string of bits required by the computer.

Such schemes may be set up in many ways. For example, in FAP, the language of the IBM's 7090 machine, the above two instructions would be written

```
CLA  200
ADD  300
```

The form of the instructions has been chosen to provide something easily written and easily recognized by the programmer, and interpretable by a not too elaborate assembly program.

Assembly languages can be quite elaborate, the peak of their rococo magnificence being reached in FAP, which has about 250 different instructions. Besides containing a great number of instructions, these languages can also have a range of ingenious devices to ease the programmer's task. The more useful of these are the labelled instruction, the cycle, and the subroutine, which we will now consider in turn.

When the programmer wishes to use a jump to some remote part of the program, it is exceedingly inconvenient to have to specify the actual address of the instruction; partly because it involves a lot of counting by the programmer, mostly because, if an extra instruction is inserted, all subsequent ones will be moved up one place and all the jump instructions must be changed. This inconvenience is avoided by allowing the programmer to label an instruction with a name, such as TEDDY, or with a number, and to refer to it thereafter by this label. For example, if the first of the two instructions above had been written

```
TEDDY CLA  200
```

it would be possible to jump to it from another part of the program by the instruction

```
TRA TEDDY
```

without knowing at which address CLA 200 was actually stored (TRA is the FAP instruction for an unconditional jump). The assembly program performs the tedious task of remembering where the instruction labelled TEDDY has been stored and seeing that the jump instruction is set up with the correct address.

In most calculations a few simple operations have to be repeated a great number of times. For example, it may be necessary to read a large number of observations from a paper tape, multiply each by a calibration constant, and store the results in successive addresses in the store. If this had to be done for 1000 observations, it would be intolerable to have to write the string of instructions a thousand times over. What is required is to run through the same instructions 1000 times, each time moving up the address in which the observation is stored. Such a set of instructions is called a cycle. A special store called an index register is provided to do the counting and to shift the addresses. This is first set to 1000 and is decreased by one each time the cycle of instructions is run through; when it gets to 0 the machine jumps to the instruction following the cycle. Index registers are useful and not very expensive devices, and counting in them is faster than counting by adding 1's to a number in a storage location, and it does not disturb the accumulator. Modern machines are provided with them: thus, EDSAC has 2, the 7090 has 3, the Control Data Corporation's 1604 has 6, and Ferranti's Atlas has 128.

It is extremely convenient for the programmer if he can divide his task into more or less self-contained pieces, and if a number of programmers are to work on the same program, it is almost essential. Such pieces of a program are called subroutines.

A subroutine may perform a simple operation such as calculating sin θ, or it may be an elaborate piece of programming which itself uses other subroutines. For example, a program for analysing observations of the position of a satellite might contain a subroutine for computing positions from measurements of a photographic plate containing images of stars and of the satellite; this subroutine might use another which found the star positions, using data from the 'Nautical Almanac'; this subroutine in turn might use others for solving spherical triangles, and these in turn would need subroutines for calculating sines and cosines. An elaborate program may contain 100 or more subroutines arranged in a hierarchy of this kind. The advantages of such a subdivision are very great. The individual subroutines can be tested separately or in small groups, and changes can be made in them without disturbing the rest of the system. It is also usually possible to use already existing subroutines for some parts of the program.

In addition to manipulating numbers in its storage, a computer must be able to take in information and give out results. Information may be read from paper tape, magnetic tape, or punched cards, or received over a land line or a radio link; for most purposes, it is transferred to magnetic tape before entering the computer. The tape provides a compact and easily handled temporary repository for the information until the computer is ready to deal with it.

With many machines the writing of programs for input and output is the most complicated and difficult part of programming. This is because of the need to provide for the appropriate action when errors occur and the necessity for allowing the tape readers to work at their own speed, which is much slower than that of the computer. It is at this point that the microsecond time scale of the computer meets the more or less human speeds of moving tapes and type bars. The programmer must prevent the computer from hurrying on to use information while it is still being read in, or from wiping out numbers before they have been printed.

Autocodes: Fortran and Algol

Writing a program in a language such as FAP is very much more convenient than directly writing the strings of bits needed by the machine, but it is still a tedious process. Everything has to be thought out in detail, and the programmer has still to write an instruction for each binary instruction required by the machine. If two numbers are to be added, the programmer must provide an instruction to bring one of them into the accumulator, first remembering to store the number previously there if it is needed again. He then writes the instruction for the addition and must consider whether it is necessary to store the sum, or whether it can be left in the accumulator; if he stores it, he must remember where he has put it and make sure that he does not put it in a storage location that already contains a number that he will need later on. The seemingly endless series of such trivial decisions is irritating, and inevitably leads to errors that may be quite difficult to discover. Even experienced programmers make one or two 'trivial' errors per 100 instructions. Another source of annoyance is the difficulty of transferring programs between different machines that have different sets of instructions and different ways of writing them. All these troubles are accentuated by the rapid expansion in the use of computers and the need to train large numbers of men and women of average intelligence to write programs that will work with reasonable efficiency and without too prolonged a hunt for errors.

It might be thought that these difficulties should be mitigated by elaboration of the machines to make their logical structure closer to that of the problems to be solved. Unfortunately the problems are very various. Some are arithmetical, some logical, some classificatory, and some linguistic, so that there is little prospect of satisfying a large proportion of users in this way. The most profitable approach has been the devising of new programming languages to express problems in terms natural to the user and to provide very sophisticated programs, called compilers, that will convert the chosen language to machine instructions. In the ideal programming language, the user thinks and writes in a way that is natural for his problem; he does not need to know where things are stored or how they are represented in the machine, and he is not concerned with accumulators, index registers, or machine instructions; all these are attended to by the compiler. The peculiarities of the machine are irrelevant, and a program can be run unchanged on any machine that has a suitable compiler. Actual compilers do not achieve this ideal. There may be peculiarities that prevent complete compatibility between a program and a machine—for example, a program may require the results to be printed in upper and lower case letters, which cannot be done on a machine that has only capitals on its printer—or a program may fail altogether because the machine does not contain as much storage as the one for which the program was written.

The most important languages are those intended primarily for the programming of arithmetic processes. The first and most widely used of these is Fortran, designed for the IBM 704 machine and its successors (currently the 7090 and 7094). Almost all machines will now accept Fortran programs.

A Fortran program [1, 2] consists of a series of statements, some of which call for arithmetic to be done, some for logical decisions, and some for input or output to the machine. Each Fortran statement produces, on the average, about five machine instructions. Arithmetic statements resemble ordinary algebraic expressions. For example, the statement

```
X=(Y-Z)/(Y+Z+2.)
```

would cause the previously determined values of Y and Z to be substituted in the expression on the right-hand side and the variable X set equal to the result. A selection of frequently needed functions is available, and the appropriate one is automatically used when its name is mentioned. Thus a function called ABSF produces the absolute value of a quantity (the value with the sign set positive), and SINF produces the sine:

```
X=ABSF(Y+SINF(Z+Y))
```

will set X equal to $Y+\sin(Z+Y)$ with its sign set positive.

A small selection of conditional jumps is provided. For example, the statement

```
IF(A) 1,2,3
```

will cause the machine to take next the statement labelled 1 if A is negative, the statement labelled 2 if it is zero, or the statement labelled 3 if it is positive. Expressions such as those on the right of the above arithmetic statements can be used in place of A, and the expression will then be evaluated and the value of the result used as the criterion for the jump.

Cycles are provided by a DO statement, for example the statement

```
DO 43 N=1,10
```

will make the machine run through all statements up to the one labelled 43, first with N=1, then with N=2, and so on up to N=10. A string of numbers can be represented by A(N) where N runs through the integers 1, 2, . . .

As an example of part of a Fortran program, suppose that a series of 100 numbers are stored as A(1), A(2), . . . , A(100), and we require to find the sine of the largest. A program to do this might be

```
  B=A(1)
  DO 1 N=2,100
  IF(A(N)-B) 1,1,3
3 B=A(N)
1 CONTINUE
  S=SINF(B)
```

In words this means 'Set B equal to the first of the 100 numbers, then try each of the remaining 99 numbers in turn, and see if it is bigger than B, and if it is, set B equal to the number'. This sequence results in B being set equal to the largest of the numbers. S is then set equal to sin B. (The CONTINUE statement has no arithmetical effect—it is included merely to give a destination for the jump from the IF statement. The 1 and 3 on the left are the labels showing the destinations of the jumps from the IF statement; the 1 also marks the end of the set of statements run through for N=2,3, . . ., 100).

Fortran has been extremely successful. It can be learnt in a few hours and gives programs that are usually as fast as those written by a skilled programmer using machine language; probably 90 per cent of all programming is now done in Fortran. Its success is also in large measure due to the existence of efficient operating systems with which it is used. The operating systems are programs that will deal with a series of jobs one after the other without intervention by the operator. Each job may contain a program of which parts are in Fortran, parts in a language such as FAP, and parts in the basic binary code; the programs needed to interpret these and load them into the machine with the correct links between them are automatically called in by the operating system when needed.

Most programs can be written in Fortran, but for elaborate logical structures, particularly non-numerical ones, such for example as a program for playing chess or for the analysis of syntax, programming in Fortran becomes tedious, owing mainly to the paucity of conditional statements. During the last four years, more powerful languages have been developed to enable the programmer to write more compactly than he can in Fortran. The best known of these is Algol, whose rules were laid down by an international committee in 1960 [3]. As a sample of the facilities provided, let us take the Algol conditional statement corresponding to the Fortran IF. In Algol one can write

if B **then** S1 **else** S2

where B is any proposition that can be true or false and S1 and S2 represent pieces of programming of any length and complexity. In Algol, words, such as **if**, that are to be recognized by the compiler as having special meanings are written in heavy type. If B is true the machine does S1, if it is false it does S2. B may be something quite simple such as $x=3$, or it may be a complicated condition such as

$$(x > (3+\sin\theta) \wedge y=0) \vee x=7\times y,$$

where $\wedge$ means 'and' and $\vee$ means 'or'. With this condition for B, the **if** statement means 'if x is greater than $(3+\sin\theta)$ and $y=0$ or if $x=7y$ do S1, otherwise do S2'. The proposition B can itself contain further **if's** nested inside it and giving a choice of propositions for testing, for example one could write

if (**if** x=0 **then** a=2 **else** b=0) **then** S1 **else** S2

meaning 'consider whether x is zero; if it is, consider whether $a=2$; if this is true, do the piece of program S1; alternatively, if x is not zero, consider if $b=0$; if it is, do S1; otherwise do S2'.

Obviously the difficulty of writing and understanding such Chinese-box-like remarks increases very rapidly with the degree to which their components are nested one within the other, and it is usually unwise for the inexperienced programmer to attempt too compact a style. There is in fact considerable difficulty in describing the language unambiguously. The Algol committee does this by the use of a symbolism specially devised for describing Algol statements. This remarkable achievement has resulted in a report which is perhaps the least readily intelligible piece of meaningful prose ever written. There is an urgent need for introductions to Algol suitable for teaching users who are not full-time programmers, and who do not need the full potential complexity of the language [4].

Some of those concerned with Algol regard it as an immutable system that has been completely defined and must not be changed. They meet the difficulties of writing efficient compilers for it, and of understanding and using it, in the spirit of lawyers or theologians glossing and defending the sacred text of the report. Not unnaturally, many users have differed from the committee on various points of detail and have proposed modifications. Many of these are concerned with limiting the programmer's freedom in the interests of providing more efficient programs for the machine. One of the most controversial of these proposals is to restrict the general requirement for all Algol subroutines to be recursive, that is to be able to use themselves. In Algol one can define functions in terms of themselves. For example, the factorial function $n!=1\times 2\ldots n$ can be defined by:

$$1!=1$$
$$n!=n(n-1)!$$

To deal with such situations complicates the compiler and, what is worse, complicates the programs produced. C. Strachey and M. V. Wilkes [5] suggest that the programmer should be required to say when a piece of programming must be able to be used recursively. The compiler can then give a more efficient program for other tasks when it is told that it need not be recursive.

Clearly there is room for differences of opinion on such matters, and there is some conflict of interest between those concerned with programming languages *per se*, who strive for the greatest generality; those interested in writing compilers, who are concerned with producing fast and compact programs and with reducing the time of compilation; and the ordinary user who is interested in having something that is easy to learn, that enables him to use programs written for other machines, and that has rules that are not changed. In practice, the swing to Algol and its variants is slow, and most programming is still done in Fortran.

The writing of compilers

The production of the first compilers was a lengthy and expensive matter. The writing of the Fortran compiler for the IBM 704 machine is said to have occupied 50 man-years. Since then more general methods have been developed, and it is now possible to produce a compiler *de novo* in about 5 man-years. It is even possible to produce programs that will write compilers, given the rules of the language to be compiled. The methods of writing a compiler are hierarchical. A few simple processes, such for example as recognizing characters, are programmed as subroutines in machine language, and these are combined to give the bare bones of a compiler. The rest is written in this restricted language, which gradually grows and is able at each stage to compile extensions of itself.

Special-purpose programs

Programming in any general programming language such as Fortran or Algol involves breaking the problem down into its arithmetical or logical elements, and expressing these in the appropriate way. There are many classes of calculation that occur so often that it is tedious to do this, but at the same time are so varied that a simple program will not deal with all of them. Examples are business account keeping, the investigation of crystal structures, and the analysis of time series. There has recently been a tendency to produce rather elaborate programs which enable such classes of problems to be expressed in special languages appropriate to themselves.

For example, a program called BOMM deals with the reduction and analysis of data that can be regarded as a set of time series [6]. In such calculations, one needs to be able to read data recorded in many different ways; to check it for gross errors; and to perform operations such as smoothing, filtering, cross correlation, and Fourier transformation. As an example of the statements of BOMM consider

```
READ TREVOR
ERROR TREVOR 1000 −1000 10 15
COSTR TREVOR PERCY 0 ·1 1
OUTPUT PERCY
PLOT PERCY
END
```

The first word in each statement is its name. The first one, READ, will read a series from the magnetic tape holding the input to the machine and will christen it TREVOR (we do not consider here how the layout of the input data is described). The statement named ERROR then causes the terms of the series TREVOR to be examined; any that lie outside the range ± 1000 or give 1st differences exceeding 10, or 2nd differences exceeding 15, are rejected as gross errors and replaced by interpolated values. COSTR then forms the cosine transform of TREVOR and calls it PERCY. The statements OUTPUT and PLOT give the terms of PERCY as a table and as a graph. Each statement makes use of hundreds of machine instructions which are themselves derived from a Fortran program.

Such a system will do nothing that could not be done by programming in Fortran or Algol. It does, however, make extensive facilities very easily available to those not wishing to indulge in elaborate programming themselves. BOMM has been mainly used in the analysis of data concerning seismology, ocean tides and waves, and the earth's magnetic and gravity fields: some of the jobs use over 300 BOMM statements. Objections to such systems are that the original programs are quite troublesome to write and never provide everything that the user needs. They also require elaborate descriptions and tend to be slower than programs written specially for a given job.

Conclusion

The capital and running costs of equipment to do a given amount of computation have been decreasing rapidly and are likely to continue to fall. The cost of programming a problem is often a large part of the total cost and may soon become the predominating cost. It is therefore of considerable importance to provide powerful programming languages that are easy to learn and to write, that provide fast and compact programs and good facilities for diagnosing errors.

It has been well said [7] that the computer can do for thought what the steam engine did for muscle. The steam engine does not do anything that could not be done by men, horses, or windmills, but it does it more cheaply and on a larger scale. It is no criticism of a steam engine to say that it cannot bath a baby or carve the Venus de Milo. Similarly a computer can relieve a man of many jobs he could do for himself but does not want to, such as account keeping and solving spherical triangles. A computer does what it is told, and it is no criticism of it to say that it cannot be told how to do the work of the Prime Minister or write a good poem. How far the computer can go in performing what are ordinarily regarded as intellectual tasks is not clear—it depends on what processes can be clearly described. The computer can carry out any mental task that can be precisely defined, and precise definition is not incompatible with the ability to learn by experience as the work proceeds. One could, for example, write a program for playing poker which would study and make use of the opponent's habits as shown in his past play. The computer gives a new and very practical meaning to Wittgenstein's well-known aphorism *Was sich überhaupt sagen läßt, läßt sich klar sagen.* That one should, perhaps, not be too supercilious about the possibilities is suggested by the following passage from *Pravda* translated on the IBM 7094 machine (quoted by permission of IBM):

> Investigation of special gas fuel was conducted by scientists in connection with investigations of other shortcomings of airplane. Full results of investigation are secret, however, along/by declaration of our scientist, lowering of airplane is a result of shortcoming in basic structure of turboreactor. Surplus heating of released gases coincides with initial lowering of force of thrust/rod and power of reactor. Once again perfection of communist technology and science ensures protection of territory and total application of weapons of progress at the expense of reconnaissances of different countries, conducting stubborn politics/policy of antagonism to peoples of Soviet Union and China.

The translation is not very polished—'lowering' is an unfortunate synonym for 'bringing down', but both the technical and the polemical parts are intelligible. To improve it does not require an improvement in machines but a better understanding of the process of translation. Things that can be dealt with by formal rules, such as the order of adjective and noun in French, or that can be exhaustively enumerated, such as the conjugation of irregular verbs, give little trouble. The difficulties arise in situations where the translated text must take account of the meaning, for example in the insertion of 'the' or 'a' in English translated from Russian, or the distinction between 'that' and 'which'. It seems likely that further improvements in machine translation will be very difficult, since it will require the machine to take into account meaning as well as syntax and grammar. The problem of improving machine translation is one of understanding languages, not of understanding programming.

References

[1] Reference Manual 709/7090 Fortran. IBM, New York 1963.

[2] McCracken, D. D. 'A guide to Fortran programming', Wiley, New York. 1961.

[3] Naur, P. *Numer. Math.*, 4, 420, 1963.

[4] Dijkstra, E. W. 'A primer of Algol 60 programming.' Academic Press, London and New York. 1962.

[5] Strachey, C. and Wilkes, M. V. *Comm. Ass. Comp. Machin.*, 4, 488, 1961.

[6] Bullard, E. C., Oglebay, F. E., Munk, W. H., and Miller, R. G. 'A user's guide to BOMM.' University of California, La Jolla. 1964.

[7] Burch, G. 'The boundless age of the computer.' *Fortune*, March 1964.

Book reviews

LABORATORY PRACTICE

Laboratory Handbook, *edited by N. L. Parr. Pp. xi + 1519. George Newnes Ltd, London. 1963. £10 10s. net.*

This is a reference book covering a wide range of topics of interest to those connected with scientific laboratories. A preliminary chapter on laboratory design, organization, and administration is followed by some 40 chapters dealing with laboratory techniques, ranging from the construction of apparatus and equipment to radioisotope techniques, chemical analysis, and testing techniques. Then follow 18 chapters on specialized laboratories. The work concludes with sections on laboratory hazards and precautions; the recording, analysis, and presentation of results; and library and information services. There is an adequate index.

With a work covering such a wide field, the depth of treatment is necessarily limited, but nevertheless there is much useful information given and in most cases the material included has been wisely selected. One of the most valuable features of the book is the bibliography and list of references at the end of each chapter. There are 47 specialist contributors, and there is no doubt that each is an authority on his subject. The work is commended as a reference book both for those taking up work in the scientific field and for established workers who need a ready reference book for techniques that they are not currently using. *M. R. Ward*

COSMOLOGY

The Nature of Comets, *by N. B. Richter (translated and revised by Arthur Beer). Pp. xli + 221. Methuen & Co. Ltd, London. 1963. 63s. net.*

The celestial apparitions known as 'comets' have throughout the ages managed to attract more attention than any other bodies in our solar system—largely because of their unusual movements. More recently, a gradual understanding of their perishable structure led to a realization that comets are also extremely abundant—although for most of the time in a non-luminous state. They are vastly more abundant than any other species of astronomical bodies. That 'there are more comets in the sky than there are fish in all waters of the Earth' was conjectured by Kepler early in the seventeenth century; although both the astronomers and ichthyologists have so far been unable to agree what these numbers actually are, the importance of the cometary phenomena for fuller understanding of the origin of the solar system is not for a minute in doubt.

Therefore, the publication of this English version of Richter's monograph on the comets should be welcomed by those engaged in cometary research and also by all astronomers concerned with the studies of the solar system. They should be indebted to Arthur Beer, the translator, for having brought the contents of this best existing critical survey of our knowledge of the astronomy and physics of cometary phenomena up to date (as of September 1962) as regards new facts and illustrations. In addition, a spirited introduction to the volume, contributed by R. A. Lyttleton, sums up the relative merits of individual theories of the origin of the comets, and should be read in conjunction with Chapter 6 of the author's own text. The whole book is well illustrated and printed. *Zdeněk Kopal*

FUNDAMENTAL PARTICLES

Fundamental Particles, *by K. Nishijima. Pp. xi + 408. W. A. Benjamin Inc., New York. 1963. $11.85.*

This is an advanced text which is up to date and systematic in presentation. As it is based on lecture notes, the style is extremely condensed, and it is hardly the type of work which could be read and understood comfortably when travelling in a train. On the other hand, it is excellent for reference and would certainly be a valuable aid to anyone carrying out theoretical research in the subject.

A considerable knowledge of field theory is assumed, but the treatment is consistent and clear and remarkably comprehensive. Although no description of experimental methods is included, proper emphasis is given throughout to the interplay of theory and experiment.

The publication of books based on lecture courses has not always proved successful, but this book does not suffer from the usual faults of hasty compilations of incompletely digested and coordinated material. To facilitate quick publication, photo-offset printing has been used, but the format is not unpleasing and the mathematics clearly presented. The book can certainly be recommended to those who are concerned with the theory of fundamental particles, and it is also of value to other theoretical physicists who maintain an interest in the most basic facts of physics and whose mathematical equipment is adequate. *H. S. W. Massey*

NUCLEAR SCIENCE

Annual Review of Nuclear Science, *edited by E. Segrè, G. Friedlander, and H. P. Noyes. Vol. 13. Pp. vi+590. Annual Reviews Inc., California. 1963. $9.*

About half of the 1963 volume in this well-established series of reviews is devoted to elementary-particle physics. In view of recent developments, few will quarrel with the editors' generous allocation of space to this branch of physics. There are authoritative articles dealing with strange particle resonant states (Dalitz), and the physics of muons and muon neutronics (Feinberg and Lederman); a comprehensive account of recent developments in the analysis of bubble chamber pictures by fast digital computers (Rosenfeld and Humphrey); and an article on pion resonances by Puppi. All these reviews are very readable and will be widely welcomed by experienced research workers and graduate students alike.

Nuclear structure research is represented by articles on coulomb excitation (Stelson and McGowan) and stripping reactions (Glendenning). Stelson and McGowan concentrate on developments that have mainly occurred during the three years preceding February 1963, and they emphasize experiments using heavy ions, now readily available with tandem accelerators. Glendenning's article is perhaps less satisfactory, particularly in those sections dealing with the stripping of more than one nucleon.

Other articles of general interest to nuclear physicists are those dealing with the ranges of protons, alpha particles and neutrons (Fano), and heavier particles (Northcliffe). There is also an account of reactions suitable for monitoring a high-energy proton beam (Cumming).

The remaining articles deal with the meteorology of fallout (Bjornerstedt and Edvarson); radiation effects on macromolecules (Hutchinson); and activation analysis (Sayre). *K. W. Allen*

COSMIC RAYS

Progress in Elementary Particle and Cosmic Ray Physics, *Vol. VII, edited by J. G. Wilson and S. A. Wouthuysen. Pp. xi+358. North-Holland Publishing Co., Amsterdam. 1963. 80s. net.*

This volume is the seventh of a series of authoritative review articles on elementary particles and cosmic rays, and is the second of the current series devoted exclusively to cosmic rays. It consists of one long article by L. I. Dorman on the geophysical and astrophysical aspects of the cosmic radiation, and is, in some respects, a sequel to the exhaustive work on cosmic-ray variations published by the same author in 1957. The present article is concerned with the determination of the variations in the primary cosmic radiation from observations on the secondary radiations at different places on the Earth's surface, and includes a particularly valuable study of the Forbush effect in relation to solar corpuscular streams. The long-term variations of the diurnal changes in cosmic rays, the apparent 22-year cycle, and the possibilities of observing galactic anisotropies are also considered. This book comes at a most opportune time in view of the starting of the International Year of the Quiet Sun.

G. D. Rochester

SCIENCE OF HIGH PRESSURES

High Pressure Physics and Chemistry, *edited by R. S. Bradley. Vol. I, pp. xi+444, 100s. net. Vol. II, pp. ix+361, 80s. net. Academic Press Inc. (London) Ltd. 1963.*

Except for a brief outline of the history of high-pressure science in this century, Volume I deals essentially with the effects of pressure on the physical properties of gases, liquids, and solids. Volume II is mainly concerned with high-pressure chemistry and contains chapters on geochemistry, gaseous equilibria, equilibrium in condensed phases, and reaction kinetics; some miscellaneous physical topics have also been incorporated.

Some of the theoretical treatments are rather condensed, but this is difficult to avoid in a work of this size. However, the sections on equations of state for solids and on gaseous equilibria deal fairly fully with the fundamental aspects of the subjects. A fair balance between practical methods and theoretical considerations has nevertheless been achieved in both volumes. Although the treatise deals principally with static pressures, one chapter on shock-waves has been included; this chapter should not fail to appeal to the reader's imagination.

The expansion of the science of high pressures in recent years calls for a treatise of this type in which chapters are written by specialists in various fields.

Taken as a whole, the work is very comprehensive; it is well illustrated and indexed, and the bibliography is extensive. It should be a valuable reference work.

E. H. Baker

CALORIMETRY

Recent Progress in Microcalorimetry, *by E. Calvet and H. Prat, translated from the French by H. A. Skinner. Pp. xii+177. Pergamon Press, Oxford. 1963. 60s. net.*

Microcalorimetry is the technique of accurate measurement of very small heat changes, as little as one-thousandth of a calorie per hour, which accompany physical, chemical, or biological processes.

Calvet and his co-workers, at the Institute of Microcalorimetry, Marseilles, have been largely responsible for the considerable progress which has been made, during the last twenty years, in the construction and use of calorimeters for these measurements. The first third of the book describes the theory of microcalorimetry, together with the construction and calibration of the calorimeter. The remainder of the book deals, in two parts, with the applications of microcalorimetry to physico-chemical processes, and to biological processes. This last part is the field in which Prat has had considerable experience.

Although much of the work has already been described in 'Experimental Thermochemistry', Vol. I, 1956, and Vol. II, 1962, it is useful to have a single, more recent monograph on microcalorimetry for those contemplating work in this field. The book is well translated.

C. T. Mortimer

LOW-TEMPERATURE PHYSICS

Low Temperature Solid State Physics, Some Selected Topics, *by H. M. Rosenburg. Pp. xvi+420. Oxford, Clarendon Press. 1963. 63s. net.*

There are several short introductory books on low-temperature physics which can be recommended to undergraduates interested in the field, and there is no shortage of specialized reviews and monographs which treat individual topics in great detail and in great depth. This book aims to fill the gap between the two, and on the whole it does so very successfully: it can be recommended to research students (and others) who wish to acquire a background knowledge of low-temperature solid-state topics, reasonably detailed but not too profound.

The author says in his preface that this is a book of simple explanations, and the claim is justified, though the simplicity is sometimes slightly specious. The first two chapters in particular have a rather slapdash air: the purist will query the first equation in the book, and will probably give up reading altogether on page 5, where wavelength and wave-vector are carelessly confused. But if he persists to the third and later chapters he will find much less to query (though, for instance, figure 4.14(a) shows a u-process, not an n-process, and even the charitable will hardly forgive figure 6.7(b)), and much to commend.

R. G. Chambers

ELECTROMAGNETIC SCATTERING

Electromagnetic Scattering; Proceedings of the Interdisciplinary Conference, Potsdam, N.Y., August 1962, *edited by Milton Kerker. Pp. xii+592. Pergamon Press Ltd, Oxford. 1963. £7 net.*

In August 1962 a conference was held at Potsdam, New York, with the object of bringing together experts in widely differing fields but with a common interest, namely electromagnetic scattering of some form or another. This volume, number 5 in the International Series of Monographs on Electromagnetic Waves, is a collection of 40 papers, mainly theoretical in nature, which were presented at the conference. About two-thirds of them are American in origin, the remainder being from various European countries. Russian work is not represented.

The book is divided into six parts of about equal length, corresponding to the six sessions of the conference, under the following headings: particle scattering; light scattering in the atmosphere and space; microwave and radiowave scattering in the atmosphere; light scattering in solution; interaction in solids and liquids; and multiple scattering. The applications of electromagnetic scattering theory are not exhausted by these headings. For instance, there are no papers on the important subject of scattering from gaseous plasmas, such as the ionosphere. Nevertheless, this collection is a very valuable one; it will be of most interest to the specialist, who will find a compendium of methods and results, both in his own and in related, though apparently diverse, subjects.

E. N. Bramley

SPECTROSCOPY

The Structure of Molecules, *by Gordon M. Barrow. Pp. xi+156. W. A. Benjamin Inc., New York and Amsterdam. 1963. $2.15 paper; $4.35 cloth.*

Some understanding of how spectroscopic measurements enable us to build up a quantitative picture of the structure of molecules is essential for the appreciation of chemistry today. The student should know how it is possible to obtain the distances between the atoms in a molecule, the angles between the bonds, and so on, from measurements on spectra. This book describes fully and clearly how this is done in language that should be intelligible to the good sixth-former. The illustrations are excellent. Electronic, vibrational, and rotational energies, and their corresponding spectra are related, and there is a section on fluorescence and phosphorescence. The student will find that the book offers him a very clear appreciation of what is happening to molecules when they absorb or emit radiation. Although not intended for the specialist, the book would form a good basis for the further study of spectroscopy.

G. Van Praagh

MASS SPECTROMETRY

Mass Spectrometry, *edited by C. A. McDowell. Pp. x+639. McGraw-Hill Book Co. Inc., New York and London. 1963. £7 15s. net.*

Professor McDowell has assembled in this volume a collection of authoritatively written accounts of different aspects of mass spectrometry with the intention of providing a guide for the novice in this field and, at the same time, a reference book for the more experienced worker. Such a dual purpose is not easy to achieve, but the book succeeds in giving a well-balanced review of the subject of mass spectrometry. Its value as a reference book may be limited as a result of the commendable decision to avoid an encyclopaedic approach.

Because of its twofold purpose, the book divides into two parts. The first contains articles on types of mass spectrometers, ion sources, ion optics, electronic techniques, and high-resolution mass spectroscopes, with a chapter on vacuum techniques that has many practical details of interest and value to workers in other fields. The second part contains chapters devoted to the consideration of some applications of mass spectrometry—chemical analysis, isotope abundance measurement, free radicals, ionization and dissociation of molecules, and ion-molecule reactions. While the treatment is not exhaustive, a good picture is provided, and one that is certain to excite interest.

The book is excellently produced, but its price may well limit its circulation to libraries and to those centres where an interest in mass spectrometry already exists.

James Bell

CHEMISTRY

Chemistry, *by T. L. Cottrell. Pp. viii+239. Oxford University Press, London. 1963. 10s. 6d. net.*

The continually widening scope of chemistry makes the development of an overall view, at an introductory level, within a single book increasingly difficult. Mention of a few of the topics covered in the present book immediately exemplifies this. It runs from electronic structure of atoms, through free energy and chemical change, and the chemistry of natural products, to modern industrial processes, the relative costs of materials, with a final chapter on the structure of chemical research.

The problem is made more difficult by the dominant influence on the teaching of elementary chemistry of the historical development of ideas. The operational definitions of concepts (for example, a pure compound) which were derived from the study of matter in bulk have now become fairly embedded in the teaching of elementary chemistry. On the other hand, these concepts are now frequently more simply defined and understood in terms of the atomic structure of matter. The author begins by treating the former type of definition with a strict regard for nicety, which means giving provisional definitions, and at a later stage in the development providing the appropriate qualifications. But treatments in terms of atomic theory are frequently limited, understandably by space considerations, to little more than outlines. The result is an unusual and slightly curious mixture of rigour, hand-out, and mile-a-minute survey. The author of the book admits that this attempt to provide a brief account at an introductory level—sixth form or first-year university—is a fairly sophisticated one.

The good student is likely to find the mixture stimulating and will probably gain a fair perspective of the basis of modern chemistry, though some teachers may think that insufficient attention has been paid to inorganic chemistry generally, and that the energetics and mechanisms of chemical change deserve more emphasis.

D. J. Millen

Annual Reports on the Progress of Chemistry for 1962. *Vol. LIX. Pp. 565+x. The Chemical Society, London. 1963. £2 net.*

This volume of Annual Reports follows the well-established pattern of earlier years in its arrangement into sections dealing with general and physical chemistry, inorganic chemistry, organic chemistry, biological chemistry, analytical chemistry, and crystallography. Only in the first of these, to which eight authors have contributed, does the presentation take the form of a review article. The special topics covered are ultra-high vacua, gas-phase oxidation, electron-spin resonance, nuclear magnetic resonance in electrolyte solutions, the equilibrium properties of mixtures, and the spectroscopy of diatomic molecules. Much of the remainder of the text is a highly condensed account of important recent publications in the general fields mentioned above, with full literature references. The ill-defined nature of the boundary between organic and biological chemistry is emphasized by the inclusion of sections on the biosynthesis of porphyrins, vitamin B_{12}, and the nicotinamide coenzymes under the latter heading. A full subject and author index is provided. Chemists throughout the world are grateful to the Chemical Society for this publication and also to those who have undertaken its compilation. It continues to occupy a unique position in the scientific literature.

H. J. Emeléus

PHYSICAL CONSTANTS

Solubilities of Inorganic and Organic Compounds, *edited by H. Stephen and T. Stephen, Vol. I, Parts 1 and 2,* Binary Systems. *Pp. 1933. Pergamon Press, Oxford. 1963. Part 1, £10 net; Part 2, £12 10s. net.*

The Russian edition of this work, published in 1962, was compiled from a literature survey made by a panel appointed by the U.S.S.R. Academy of Sciences. The English edition includes 47 tables compiled later. Tables 1 to 1694 present solubilities of substances in water, and tables 1695 to 6539 present solubilities in non-aqueous solvents. There is a list (not in English alphabetical order) of 208 references to Soviet literature and an alphabetical list of 1746 non-Soviet references followed by a second set of tables giving the compositions of stable solid phases.

A formula index is divided into two parts. The first is said to be arranged in order of the periodic classification of the elements but in fact starts with the entries H_2, He, N_2, O_2, N_2+O_2, O_3, Ne, Ar, and continues in this odd manner. There is also an index of compounds; it is claimed that the I.U.P.A.C. nomenclature has been used throughout. Some systems appear in one index but not in the other.

Solubilities are usually expressed either in wt. %, vol. %, mole %, or g/l, but the units employed for gas solubilities are varied and in some instances are incompletely defined. In general, only one value is given for each solubility, and thus a different approach has been adopted from that used in Seidell's well-known book in which averaged values are listed and discordant experimental values are recorded.

A few tests quickly satisfied the reviewer that the abstraction from the literature is far from complete. This volume illustrates the difficulties of the retrieval of data from modern scientific literature. The importance of solubilities to chemists is so great that it is to be hoped that this subject will receive attention under some future project, such as the N.B.S. National Standard Reference Program.

E. F. G. Herington

POLAROGRAPHY

Modern Polarographic Methods, *by Helmut Schmidt and Mark von Stackelberg, translated by R. E. W. Maddison. Pp. viii+99. Academic Press, New York and London. 1963. 44s. net.*

This short monograph amounts to a high-level discourse on the relative merits of several recent instrumental developments of polarography. It will be appreciated by practising polarographers, by those who have taken a graduate course in electroanalytical methods, but by few others. It assumes, for example, a thorough knowledge of ordinary d.c. polarography, as well as a basic familiarity with various special solutions of Fick's second law. No attempt is made to consider specific systems, and only passing reference is made to electronic details.

Nevertheless, in spite of these restrictions, it remains a very satisfying book. It is concerned with general principles and with their practical implementation. It is based on a deep understanding and practical experience of such sophisticated methods as cathode-ray polarography, pulse polarography, radio-frequency polarography, and tensammetric and rectification techniques. Anyone using or contemplating the use of any of these methods could read this book with considerable profit. There is clearly much fundamental and exciting research still to be done in this field, and this book would be an excellent starting point. It is likely also that all forms of polarography will continue to prosper, since the purely electrical read-in and read-out of the method is particularly suitable for automatic analysis.

The translation from the original German text is a literal one and is entirely adequate except for the oft-repeated misrepresentation of 'faradaic' as 'faradic'.

G. J. Hills

BORON HYDRIDES

Boron Hydrides, *by William N. Lipscomb. Pp. ix+275. W. A. Benjamin Inc., New York and Amsterdam. 1963. $15.40.*

One of the outstanding advances in chemistry during the last decade has been the elucidation of the structure and bonding of the boron hydrides and their derivatives. Professor Lipscomb's book gives an authoritative and very readable account of this work.

The main text, which is beautifully illustrated, occupies some 200 pages and consists of five chapters. The first two present the detailed geometry of individual molecules and the application of 3-centre bond theory to the boron hydrides. The third chapter deals with the molecular orbital description of polyhedral molecules and ions; it compares theory with experiment and makes several valuable predictions. The chapter on proton and boron n.m.r. spectra has a valuable collection of graphical and numerical data.

The final chapter is a tabular summary of chemical reactions classified under general reaction types. This section will be too condensed to be helpful to those not already familiar with the chemistry of the boron hydrides, but it forms a ready source of references to these reactions. Mention of properties which are useful for identification, such as melting point, vapour pressure, and infrared spectra, has been omitted entirely.

There are 350 references, of which more than half are concerned with work done in the last five years.

N. N. Greenwood

CRYSTALLOGRAPHY

The Chemistry of Imperfect Crystals, *by F. A. Kröger. Pp. xvi+1039. North-Holland Publishing Co., Amsterdam. 1964. £11 net.*

The preparation and study of crystals has become an important part of current physical research. Too often in the past the influence of structural and compositional imperfections on the interpretation of the results of such studies has been ignored, although the use of increasingly refined techniques has emphasized the importance of the defect crystalline state. This work deals specifically with the problems of the preparation and the physical chemistry of imperfect materials of the non-metallic, inorganic type. The author is eminently suited to the preparation of such a book, since he, and the laboratories with which he is associated, are major contributors to this field.

Although not the first volume dealing with the subject, it is certainly the most comprehensive. The second and largest part of the book consists of a detailed description of the solid state, and the development of a thermodynamic approach to solid state equilibria; a variety of materials are considered in some detail. The first section deals adequately with the topics of crystal growth, purification, and doping, while the final section gives examples of the application of crystal imperfection theory to various solid-state reactions.

The very extensive references given, together with the numerous tables summarizing published work and data, will make this an important reference book. It will be of particular value for that currently developing hybrid, the 'materials' scientist.

E. A. D. White

PHYSICAL ORGANIC CHEMISTRY

Advances in Physical Organic Chemistry, *edited by V. Gold, Vol. I. Pp. xv+443. Academic Press, London and New York. 1963. 90s. net.*

While many chemists will question the need for yet another 'Advances' series, those with physical-organic interests will be more ready to welcome it, for hitherto reviews in this field have been inconveniently widely scattered. This series should make it easier to have an adequate selection at hand.

L. L. Schaleger and F. A. Long begin Volume I with a timely review on the relationship between activation entropies and reaction mechanisms, and this is followed by an excellent summary by L. M. Stock and H. C. Brown of the work that led to a quantitative treatment of substituent effects in aromatic substitution. A. I. Shatenshtein contributes on hydrogen-isotope exchange reactions in liquid ammonia.

These three chapters on reactions are followed by three concerned with structures. G. Ferguson and J. M. Robertson write on planar and non-planar aromatic systems; J.C.D. Brand and D. G. Williamson on the structures of electronically excited molecules; and M. C. R. Symons provides a particularly valuable account of the use of electron-spin resonance for the identification of organic free radicals.

Both Dr Gold and the publishers are to be commended for the speedy publication of this volume. It sets a standard that one hopes to see maintained in its successors.

J. G. Watkinson

ALIPHATIC FLUORINE COMPOUNDS

Chemie und Technologie aliphatischer fluororganischer Verbindungen, *by Dieter Osteroth. Pp. viii+195. Ferdinand Enke Verlag, Stuttgart. 1964. Bound, DM54.*

This book contains five chapters which describe fluorinated paraffins, olefines and acetylenes; fluorinated alcohols and ethers; fluorinated aldehydes and ketones; fluorinated acids; and fluorinated nitrogen- and sulphur-containing compounds.

The first chapter brings a feeling of disappointment because of apparent omissions and a preoccupation with the patent literature to the exclusion of the less prejudiced normal sources. In the subsequent chapters, however, most of the omissions, such as the preparation of olefines by heating sodium salts and the preparation of iodides from silver salts and iodine, are rectified, and it becomes evident that the high proportion of references from the patent literature bring in some useful examples of reactions which do not appear elsewhere. Fortunately, the excellent index makes it easier to find reactions which have been misplaced or not duplicated in the appropriate chapter.

The use, as acylating agents, of the mixed anhydrides of trifluoroacetic acid with other acids is not mentioned, nor is permanganate in acetone solution as an oxidizing agent for olefines; telomerization reactions should have been dealt with more fully, and the additions to olefines would be better appreciated if split up into electrophilic, nucleophilic, and free radical reactions; but this is a very short list of complaints to arise from a book which covers almost completely the known reactions in the fields delineated.

It is recommended to all chemists as an introduction to the chemistry of fluorinated organic compounds and, in particular, to those about to become involved in research work using these compounds. A second volume describing fluorinated aromatic and heterocyclic compounds is promised.

W. K. R. Musgrave

GEOPHYSICS

Geophysics: The Earth's Environment. (Lectures delivered at Les Houches during the 1962 Session of the Summer School of Theoretical Physics, University of Grenoble), *edited by C. DeWitt, J. Hieblot, and A. Lebeau. Pp. xiv+624. Gordon and Breach, Science Publishers, New York and London. 1963. Paper, $8.50; cloth, $10.50.*

The 1962 session of the summer school of theoretical physics in the University of Grenoble was devoted to a study of the Earth's upper atmosphere and related subjects. The lectures given are collected together in this book. They dealt with: the theory of plasmas (Denisse); magneto-ionic theory (Budden); charged particles in the Earth's magnetic field (Chamberlain); scattering of waves by plasmas (Benoit); composition of the upper atmosphere (Nicolet); the ionosphere (MacDonald); airglow and the aurora (Barbier); solar plasma, geomagnetism, and aurora (Chapman); the exosphere (Dungey); whistlers (Gallet); and atmospheric movements (MacDonald).

As would be expected from these authors, the treatment is scholarly and authoritative. The material has been selected with care and presented in a particularly clear and concise manner. Fundamentals are stressed, and the newer parts of the subject are discussed without too much elaborate detail.

The publication of the book is particularly opportune at a time when there is such an overwhelming flood of papers concerning the upper atmosphere and the Earth's environment. Although the subject is advancing with unusual rapidity, so much of this book is concerned with fundamentals that it will be a long time before it becomes seriously out of date. It will provide a valuable guide both to those new to the subject and to specialized workers who wish to bring themselves up to date with the work of their fellow specialists.

J. A. Ratcliffe

CLOUD PHYSICS

Thermodynamics of Clouds, *by Louis Dufour and Raymond Defay, translated by M. Smyth and A. Beer. Pp. xiii+255.* International Geophysics Series, *Vol. VI. Academic Press, London and New York. 1963. 80s. net.*

The reviewer must confess to having little sympathy for this book, which has a misleading title and appears to serve no very useful purpose. It is emphatically not about the thermodynamics of clouds in the sense of discussing phase and energy changes in moving masses of cloudy air. It deals only with very limited (equilibrium) aspects of the existence of (single) droplets and crystals in clouds.

The first seven of the thirteen chapters are concerned with basic thermodynamics, which is adequately covered in other books, such as Guggenheim's 'Thermodynamics'. The next three chapters examine, in great detail, the stability of that equilibrium of a single droplet, ice crystal, and combination of the two in an atmosphere of constant temperature and vapour pressure. Such academic matters as the compressibility of the drop, the influence of its curvature on surface tension, equilibrium temperature, and latent heat of vaporization are investigated in detail, but these effects are shown to be of negligible importance in cloud physics problems. The last three chapters are concerned with the homogeneous nucleation of water droplets and ice crystals along the lines followed by Volmer, Becker and Döring, and Turnbull and Fisher, but the authors do not emphasize sufficiently the limitations of the theory and of the experimental data.

We are concerned, in natural clouds, not with isolated drops and crystals in equilibrium conditions but with whole populations of growing (or evaporating) particles. From this point of view the aspects of the problem treated in this book are quite trivial and can be treated satisfactorily in a much shorter space.

B. J. Mason

INDUSTRIAL MICROBIOLOGY

Biochemistry of Industrial Micro-organisms, *edited by C. Rainbow and A. H. Rose. Pp. xix+708. Academic Press, London and New York. 1963. 147s. net.*

This book comprises a collection of articles on the major classes of microbial products of actual or potential industrial interest. Also included are chapters on pathways of carbohydrate fermentation, effluent disposal, and microbial spoilage. All the major industrial fermentations are included, except for the food fermentations, which seems an arbitrary omission. The emphasis is on the chemical nature of the products and their biosynthesis rather than on the physiology of the fermentations. A different approach is taken in the article on microbial food, which discusses the use of micro-organisms for food purposes. Also, in the chapters on baker's yeast and sewage treatment, the discussion is more on the biological than the chemical aspects. However, these different approaches reflect the main progress so far made in the different fields. The scope of the articles varies considerably, from that of the comprehensive review, for instance, of steroid transformations, gibberellins, and enzymes and vitamins, to outlines of the main features of the subjects, as with the antibiotics and microbial spoilage. The articles are authoritative and show the latest position in a wide field of industrial microbiology. Though it is a miscellany rather than a synthesis of knowledge, this book makes interesting and stimulating reading, and will prove a valuable work of reference.

S. J. Pirt

PROTEIN STRUCTURE

Ultrastructure of Protein Fibers, *edited by Rubin Borasky. Pp. xvii+185. Academic Press, New York and London. 1963. 72s. net.*

This book consists essentially of the papers given in a three-hour symposium at the annual meeting of the Electron Microscope Society of America at Pittsburgh in August 1961, and this origin explains two of its features, the concentration on a limited number of specific aspects and the superb micrographs, optical and electron, which are so beautifully reproduced. A survey of keratin and flagella structure takes up most of the book; the other chapters are on physical, chemical and mechanical properties (chiefly of silk), on collagen (chiefly swollen), and on muscle (chiefly 'catch' muscles). As the editor says, the ultrastructure is interesting not only for its own sake, but also because of the biological role of the fibres and their technological importance. However, technology is rarely mentioned except in the chapter on mechanical properties—and even here the use of grams per denier as a unit of stress may confuse the enquiring biologist. On the whole this book lacks the cohesion given by a single author or an executive editor, but, with its admirable reference lists and illustrations, it would be a first-rate member of a series of books describing recent advances.

N. Peacock

NUCLEIC ACIDS

Essays on Nucleic Acids, *by Erwin Chargaff. Pp. xii+211. Elsevier Publishing Co., Amsterdam. 1963. 45s. net.*

When the reviewer states that after reading this book for the present purpose he intends to read it again for his own pleasure, some eyebrows may be lifted—but no matter: that is what he hopes to do. The work is not a textbook of the usual kind for, though it has a serious educational aspect, it is also a piece of literature enlivened with many flashes of wit. The racy style, however, cannot hide the deep philosophic insight which characterizes these essays.

The author's appreciation of modern progress can fairly be described as critical. In effect he utters a much-needed protest against the facile dogmatism of some enthusiastic theorists in the field of nucleic acids and their biological significance.

Although the book is not a guide to the organic chemistry of nucleic acids, it gets down to the bench in chapters on the isolation and purification of the nucleic acids from different sources. Other chapters deal with the properties of nucleoproteins and naturally the part played by DNA and RNA in the transmission and reception of messages of biological importance. Even the technical sections are eminently readable, and should be carefully studied by all those directly concerned with the subject. But it is also for the general reader, especially if he is interested in the logic of scientific development and the psychology of scientists. *R. Robinson*

FOOD SCIENCE

Introduction to the Biochemistry of Foods, *by J. B. S. Braverman. Pp. xv+336. Elsevier Publishing Co., Amsterdam. 1963. 70s. net.*

As the author states in his preface, foods are complex biochemical systems, many of them with a 'life' of only a few days, so that the education of a scientist who will work with food and the food industry must include a course in the biochemical aspects of his subject. The older tradition was for chemistry to be considered as the basic discipline, as though foods were stable inorganic commodities to be analysed and characterized by the agricultural chemist. On the other hand, the writer of the ordinary textbook of biochemistry has to be selective in his examples, and naturally pays most attention to aspects appealing to the medical students who form the majority of his readers. This type of book does not connect directly with the rest of the training of the food scientist. Professor Braverman has escaped entirely from this tradition and deals directly with the biochemistry of foods. In addition to sections on enzyme kinetics, glycolysis, and protein structure, for example, he also includes chapters on enzymatic and non-enzymatic browning and the metabolism of pectic substances in vegetable materials, and he discusses the mechanisms catalysing the development of rancidity in fat, cloudiness in beer, and artificial ripening of fruit. There is a bias towards the author's special interest in citrus fruit, and it would probably be difficult for a student to follow the argument at some points without reference to some more general biochemical texts, but the book has a definite place of its own. *K. J. Carpenter*

BOTANY

The Flowering Process, *by Frank B. Salisbury. Pp. xii+234. Pergamon Press, Oxford. 1963. 50s. net.*

The control of flowering is a difficult subject to review satisfactorily in the present state of knowledge. In the past fifty years an immense amount of factual information has been collected by observation and from physiological experiments, but the results have not yet been brought into relation with one another by any acceptable generalization. In order merely to classify the effects of day-length and temperature on flowering, the author is obliged to set up forty-eight categories and says that the list has only one real value, to indicate the fantastic diversity of response among the many species which have so far been studied. He predicts the need of even more complex categories in the future.

It is therefore a notable achievement to have produced an account that is both interesting and readily readable and which covers a great deal of the complicated story. Much condensation and selection have been necessary. So as not to lose his reader in the mass of conflicting and often unrelated facts, the author has avowedly concentrated mainly on certain aspects of the physiology of flowering in particular species well studied by himself and other workers. It would be unwise to quarrel with a decision which has led to a book which will be a great help to the hard-pressed student. *W. O. James*

POGONOPHORA

Pogonophora, *by A. V. Ivanov (translated from the Russian and edited by D. B. Carlisle, with additional material by Eve C. Southward). Pp. xvi+479. Academic Press. London. 1963. 90s. net.*

The first report of a pogonophoran was published in 1914, but attracted little notice. Not surprisingly, it proved difficult to interpret animals that may be 600 times as long as they are broad, and that, against all expectations, have completely lost their mouth and alimentary canal. Thus it is only within the last few years that these marine forms have been unequivocally characterized as a new phylum, probably related to the vertebrate line.

Recent advances owe much to Ivanov's studies of material collected by R. V. Vityaz, so that in this monograph he is able to provide an authoritative account of the group. In part a revision and expansion of the section he contributed to Volume v of Grassé's *Traité de Zoologie*, it includes also a survey of some 70 species, an indication of the progress made since 1952, when only three had been recognized.

The author's French publishers have made available many of his earlier illustrations, including some coloured stereograms. Carlisle has prepared a lucid translation, enlivened with apt personal comments. Dr Southward contributes new material and some colour photographs of living specimens. In short, justice has been done to an important zoological discovery. It is a pity that there is no index to the coloured plates. *E. J. W. Barrington*

ATHEROSCLEROSIS

Atherosclerosis and its Origin, *edited by Maurice Sandler and Geoffrey H. Bourne. Pp. xiii+570. Academic Press, New York and London. 1963. 157s. net.*

This book has the important aim of bringing together the information from different disciplines which is relevant to an understanding of the development of atherosclerotic lesions in arteries. The thirteen chapters, written by experts in various aspects of the subject, cover a wide range of topics, including the morphology and chemistry of arteries; lipid metabolism; and the effects of diet and hormones in man and experimental animals. Special mention may be made of the contribution by R. C. Buck on electron microscopy of arterial tissue; the detailed survey by S. Lindsay and I. L. Chaikoff of arterial lesions in birds and mammals; and the general account of lipid metabolism by L. Swell and C. R. Treadwell which emphasizes recent information on the fatty acid composition of lipid fractions in blood and arterial wall.

Workers in this field of research will find much of interest in the individual chapters. Some lack of balance in the treatment of the subject as a whole is suggested by the considerable attention given to the metabolism of the arterial wall in man and experimental animals, while the role of mural thrombosis in the development of human lesions receives brief mention. Since there is no agreed single cause in this disease, it is not surprising to find differences of opinion between the contributors on the significance and priorities of the factors concerned in the development of lesions.

J. E. French

CANCER RESEARCH

Viruses, Nucleic Acids, and Cancer; a collection of papers presented at the 17th Annual Symposium on Fundamental Cancer Research 1963. *Pp. xiii+659. The Williams and Wilkins Co., Baltimore. 1963. $16.*

The papers read at this symposium, dedicated to the memory of Francisco Duran-Reynals, are published in eight sections. The first three are arranged so as to give the reader a clear picture of the latest observations on the structure of viruses, their replication within infected cells, and the nature of the nucleic acids involved. There then follow a number of papers on the avian neoplastic viruses and on other oncogenic viruses that may need a co-carcinogen to make them effective. The role of interferon and the interrelationship of viruses are discussed before an account of the important recent work on the production of malignant tumours in young hamsters by the injection of human type 12 adenovirus. This raises the question of the possible oncogenic action of adenoviruses in man and, since they provoke a good antibody response, of the eventual use of prophylactic vaccines. Papers on lymphoma of the jaw in well-defined areas of Africa, and on the polyoma virus, complete a work of outstanding value to which contributions are made by many of the leading workers in this field.

The book is profusely and beautifully illustrated, the papers arranged in logical sequence so as to be most readable, and the bibliography extensive.

T. F. Hewer

BIOLOGICAL EFFECTS OF RADIATION

Cellular Basis and Aetiology of Late Somatic Effects of Ionizing Radiation, *edited by R. J. C. Harris. Pp. xii+359. Academic Press, London and New York, 1963. 80s. net.*

The symposium described in this volume dealt with the mechanisms which are responsible for the late somatic effects of ionizing radiation. Emphasis was placed on the discussion of the nature of the early biological changes which lead eventually to the appearance of cancer, leukaemia, or other delayed effects.

It is remarkable how much is known about radiobiology as a whole and how little is firmly known about the mechanisms for the induction of long-term effects. In this symposium, a very distinguished list of participants discussed some relevant questions. For instance, does ionizing radiation produce long-term effects through somatic mutations or through chromosome breaks? Is the activation of latent viruses important, or is it hormones that are important? What is the role of morphological changes produced by radiation, such as blood vessel damage? Are changes in connective tissue important? Mayneord perhaps summed up the dilemma by quoting Francis Bacon, who wrote: 'Touching the length or shortness of life in beasts, the Knowledge which may be had is slender, the Observation negligent, the Tradition fabulous'.

This must have been a very stimulating conference for the participants, and it affords fascinating reading. It will no doubt be read, or dipped into, by all who are seriously interested in radiobiology. Those who make calculations about long-term effects of various doses of radiation in man should certainly read it in case they take their calculations too seriously.

I. S. Eve

METALLURGY

Metallic Solid Solutions, a Symposium on their Electronic and Atomic Structure, *edited by J. Friedel and A. Guinier. Pp. xi+645. W. A. Benjamin Inc., New York and Amsterdam. 1963. $19.75.*

The main features of the electron theory of pure metals are comparatively well understood. When, however, a solid solution is formed by the substitution of solute atoms for those of the solvent, the accurate description and visualization of the electronic and atomic structure is a matter of considerable difficulty. This problem was the subject of an international symposium at Orsay in July 1962, and the book under review is a collection, in English, of 52 of the papers presented. All are relevant, and though they vary in length and weight, they constitute a most valuable survey of modern thought in this complex field. It is impracticable to summarize the content of the papers, which vary from the purely mathematical to the practical. However, the scope may be illustrated by reference to the discussion of the concept of virtual bound states by Friedel, and to the work of Blandin and Deplante, which demonstrates the existence of oscillations in the electron density and electrical potential around impurity atoms. Such fluctuations affect solute segregation, with which other papers deal in detail.

The volume is well produced and illustrated. As the work of reference which it will undoubtedly become, it suffers from lack of indices and continuous pagination, but remains essential reading for all concerned with alloy structures.

G. V. Raynor

ARCHAEOLOGY AND SCIENCE

Archaeology and the Microscope, *by Leo Biek. Pp. 287. Lutterworth Press, London. 1963. 45s. net.*

In order not to be misled by the title it is important to realize at the outset that Mr Biek's microscope is a symbolic one which stands for every kind of instrumentation and scientific technique. Indeed, we are constantly reminded throughout the book how richly rewarding it can be to discuss archaeological finds with scientists engaged on the study of contemporary material, be it iron, bronze, ceramic, wood, or leather. In most cases the result is that specialist knowledge and analytical facilities are brought to bear on a problem which would otherwise have remained with the single-handed general practitioner in a laboratory of limited resources. The book is an account of the author's experiences with a considerable variety of archaeological material over a number of years at the Ancient Monuments Laboratory of the Ministry of Public Buildings and Works; by the very nature of the subject matter, it should appeal to the general reader and in particular to scientists. (Archaeologists will presumably prefer the specialized journals as sources.) Readers must, however, be prepared for a text which is interspersed with the author's philosophical musings on the significance of these activities; in places this has produced some oddities such as the air-earth-fire-water diagram on p. 173 and, occasionally, some obscure passages.

S. Rees Jones

SCIENTIFIC STUDY OF WORKS OF ART

Institut Royal du Patrimoine Artistique, Bulletin VI, 1963. *Pp. 286. Institut Royal du Patrimoine Artistique, Brussels. 250 B.Frs. for the annual copy.*

This is the sixth member in the series of yearly volumes issued by the Institute in Brussels, which has for long been famous for its work in the scientific study and treatment of old works of art. The high standard we have learnt to expect is fully maintained. To the scientist, there are several contributions of major interest.

One concerns the chemistry of the binding-media of old pictures, with particular reference to the diterpene resins and their identification. In addition to spectrographic studies, nuclear magnetic resonance has been invoked. Some of the curves reproduced for the former lack definition of the ordinate (presumably percentage transmission), and they would all be the better for the inclusion of the effective slit-width, but otherwise they are excellent. Another chapter deals with the stonework of the Town Hall at Oudenarde (1526–36). Photomicrographs and analyses reveal the principal alterations since this famous building was erected.

Of particular value is the supplement by Dr Paul Coremans describing the course—lasting two years—which the Institute conducts, on most generous lines. Twelve students a year are accepted. The training they receive is a major contribution to the preservation of our heritage. *Ian Rawlins*

HISTORY OF SCIENCE

The Natural Philosopher, *Vol. I, edited by Daniel E. Gershenson and Daniel A. Greenberg. Pp. 155. Blaisdell Publishing Co., New York. 1963. $1.95 paper bound, $2.95 cloth bound.*

This volume, the first of a series to be devoted to the history of physics and to the influence of physics on human thought and affairs through the ages, contains four papers, three of which deal with the twentieth century and can be recommended to scientists generally and not only to historians of science.

In a long and thoroughly documented paper, R. E. Beardsley reviews knowledge of the effects of radiation on man, surveys the history of radiation control, and discusses some ethical problems raised by the use of radiation in medicine and industry, as well as war. He shows that even today radiation is often used in circumstances that are far from safe, and his paper deserves to be widely read.

M. J. Klein throws new light on Planck's gradual acceptance between 1901 and 1906 of the statistical interpretation of entropy, and on his developing insight into problems of the quantum theory. There is also an interesting transcription by Samuel Devons of some personal reminiscences of Rutherford, recorded on tape by William Kay, his research assistant and laboratory steward at Manchester from 1906 to 1919. The remaining contribution, a new translation, with a few notes, of the preface to Lavoisier's *Traité élémentaire de chimie* (1789), seems out of place in a publication devoted to the history of physics. *W. A. Smeaton*

Storia delle Scienze, *edited by Nicola Abbagnano. Vol. I, by Nicola Abbagnano, Giorgio Abetti, Roberto Almagià, and Ludovico Geymonat, Pp. xii+662. 10 000 lire; Vol. II, by Mario Gliozzi and Michele Giua. Pp. xvii+802. 12 000 lire; Vol. III, Pt. I, by Giuseppe Montalenti. Pp. vii+620. 18 000 lire; Vol. III, Pt. II, by Angiola Massucco Costa and Franco Ferrarotti. Pp. vii+448. 18 000 lire the two parts. 1962. Unione Tipografico-Editrice Torinese, Turin.*

This is a general history of science, intended for, say, a university student with some grasp of science. The first volume starts with an introduction to the whole series by its editor, Nicola Abbagnano, which discusses the value of a history of science, and his reasons for choosing to bring together histories of the individual sciences rather than produce a chronological survey of the progress of all scientific effort. The volume also deals, classically and not exhaustively, with the philosophy of science.

The main text contains a history of astronomy by Giorgio Abetti, one of geography by Roberto Almagià, and one of mathematics by Ludovico Geymonat. The second volume is devoted to a history of physics by Mario Gliozzi, and one of chemistry by Michele Giua; in the first book of the third volume Giuseppe Montalenti writes about the history of biology and medicine. The second book of volume III deals with the history of sciences that are not always included in works such as this—psychology by Angiola Massucco Costa, and sociology by Franco Ferrarotti.

The text has to be treated with some caution, as it is sometimes guilty of a misleading lack of historical sense. Thus Cavendish is said (page 567 of Volume II) to have confirmed that air is a mixture of oxygen and nitrogen and to have discovered that water is formed from air and hydrogen. We, with hindsight, can say that this is what he did, but he certainly would not have described his results in this way. As the next page tells us, he remained an adherent of the phlogiston theory all his life.

The book is lavishly produced and illustrated, and many of the illustrations are not the ones familiar from other histories. The text, too, is detailed, and contains notes on many minor characters who barely achieve mention in shorter histories. *W. A. Osman*

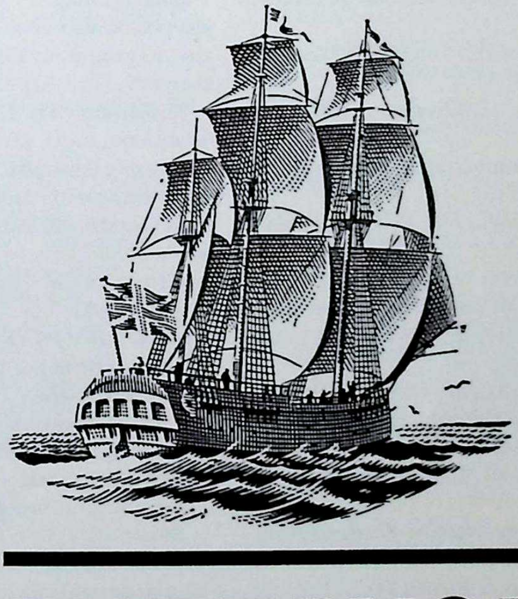
The Lunar Society of Birmingham, a Social History of Provincial Science and Industry in Eighteenth-century England, *by Robert E. Schofield. Pp. x+491. Clarendon Press: Oxford University Press, London. 1963. 70s. net.*

This work becomes at once the authority on its subject, but it does not merely conclude the long line of studies of a fascinating group. It is likely to promote new examinations of the conduct of scientific work in eighteenth-century England, because Dr Schofield insists on seeing the Lunar Society in so wide a setting that its nature and influence take on a new character.

Many writers of the history of science still cling to the idea of the 'amateur' as the typical scientific investigator before the growth of organized professionalism in the nineteenth century. Dr Schofield adopts a consistently different view. The science of the Lunar Society is seen as being no more an isolated activity than any of the other pursuits which filled the lives of Boulton, Watt, Priestley, Wedgewood, and the others. He takes the history of the Society from before its birth until after its death, with a detailed account of the activities of each of its principal members. The general picture is one of a vigorous provincial life, not dependent on London, dissenting in more than a religious sense from accepted standards; of men who were fostering the Industrial Revolution by combining scientific insight, technical resourcefulness and commercial acumen; who, in so doing, brought about changes over a wide range, from the relation of the classes to the law of the land.

It is no longer enough to recount the history of science as a succession of ideas. This book will be a strong stimulus to those who see it rather as a history of the deeds of men. *Frank Greenaway*

2/13



ENDEAVOUR

A review of the progress of science, published in five languages

VOLUME XXIII 1964

Author index

BADDILEY, J. Teichoic acids and the bacterial cell wall. 33.
BARRER, R. M. Molecular sieves. 122.
BARTLETT, N. The chemistry of the noble gases. 3.
BEGUIN, C. P., KANAAN, A. S., and MARGRAVE, J. L. Plasma chemistry. 55.
BENTLEY, K. W. The relief of pain—the search for the ideal analgesic. 97.
BRANDON, D. G. Field-ion microscopy. 90.
BROWN, R. J. E. and JOHNSTON, G. H. Permafrost and related engineering problems. 66.
BULLARD, SIR EDWARD. The language of machines. 160.
CHERRY, COLIN. The communication of information. 13.
CHILDS, E. C. The movement of soil water. 81.
EDITORIAL
  Books for students in developing countries. 114.
  Frederick Soddy and the concept of isotopes. 54.
  Galileo Galilei (1564–1642). 2.
GORDON, J. E. Whiskers. 8.
GOTTLIEB, DAVID. Germination of fungus spores. 85.
GRAY, E. G. Electron microscopy of the cell surface. 61.
HOSKIN, MICHAEL. William Herschel and sidereal astronomy. 18.
JOHNSTON, G. H. *See* BROWN, R. J. E. and JOHNSTON, G. H.
KANAAN, A. S. *See* BEGUIN, C. P., KANAAN, A. S., and MARGRAVE, J. L.
KUNZLER, J. E. Superconducting magnets. 115.
LONCAREVIC, B. D. Geophysical studies in the Indian Ocean. 43.
MARGRAVE, J. L. *See* BEGUIN, C. P., KANAAN, A. S., and MARGRAVE, J. L.
MASON, B. J. The collision, coalescence, and disruption of drops. 136.
MOON, P. B. and O'CONNOR, D. A. Some applications of the Mössbauer effect. 131.
MORTIMER, C. T. and SPRINGALL, H. D. Modern thermochemistry. 22.
O'CONNOR, D. A. *See* MOON, P. B. and O'CONNOR, D. A.
PRESTON, R. D. Structural plant polysaccharides. 153.
PURNELL, J. H. Physical measurement by gas chromatography. 142.
ROBIN, G. DE Q. Glaciology. 102.
SPRINGALL, H. D. *See* MORTIMER, C. T. and SPRINGALL, H. D.
SUTCLIFFE, R. C. Weather forecasting by electronic calculation. 27.
SUTHERLAND, N. S. The learning of discriminations by animals. 148.
THOMPSON, W. B. Magnetohydrodynamics. 73.
WEITZ, B. Feeding habits of tsetse flies. 38.

Book reviews

ABBAGNANO, NICOLA (General Editor). Storia delle Scienze, Vols. I–III. 172.
ABERCROMBIE, M. and BRACHET, JEAN (Editors). Advances in Morphogenesis, Vol. II. 110.
AMOS, BERNARD and KOPROWSKI, HILARY (Editors). Cell-bound Antibodies. 109.
Annual Reports on the Progress of Chemistry for 1962. Vol. LIX. 167.
BARNES, HAROLD (Editor). Oceanography and Marine Biology, an Annual Review, Vol. I. 111.
BARRINGTON, E. J. W. An Introduction to General and Comparative Endocrinology. 51.
BARROW, GORDON, M. The Structure of Molecules. 166.
BAUR, FRANZ. Grosswetterkunde und langfristige Witterungsvorhersage (Large-scale Meteorology and Long-range Forecasting of Weather Character). 109.
BEAMENT, J. W. L., TREHERNE, J. E., and WIGGLESWORTH, V. B. (Editors). Advances in Insect Physiology, Vol. I. 51.
BIEK, LEO. Archaeology and the Microscope. 171.
Biographical Memoirs of Fellows of the Royal Society, Vol. IX. 112.
BORASKY, RUBIN (Editor). Ultrastructure of Protein Fibers. 169.
BOS, L. Symptoms of Virus Diseases in Plants. 111.
BRADLEY, R. S. (Editor). High Pressure Physics and Chemistry, Vols. I and II. 166.
BRAVERMAN, J. B. S. Introduction to the Biochemistry of Foods. 170.
British Pharmacopoeia 1963, and British Pharmaceutical Codex 1963. 111.
CALVET, E. and PRAT, H. Recent Progress in Microcalorimetry (translated from the French by Skinner, H. A.). 166.
CAMERON, A. G. W. (Editor). Interstellar Communication, a Collection of Reprints and Original Contributions. 107.
CHADWICK, SIR JAMES (Under the scientific direction of), The Collected Papers of Lord Rutherford of Nelson, Vol. II. 112.
CHALMERS, BRUCE. Energy. 48.
CHARGAFF, ERWIN. Essays on Nucleic Acids. 170.
CHÊNE, MARCEL. Chimie appliquée à l'analyse des matériaux et produits cellulosiques (végétaux, pâtes cellulosiques, papiers). 109.
CLAYSON, DAVID B. Chemical Carcinogenesis. 50.
COTTRELL, T. L. Chemistry. 167.
DEWITT, C., HIEBLOT, J., and LEBEAU, A. (Editors). Geophysics: The Earth's Environment. 169.
DJERASSI, CARL (Editor). Steroid Reactions; An Outline for Organic Chemists. 49.
DUFOUR, LOUIS and DEFAY, RAYMOND. Thermodynamics of Clouds (translated by Smyth, M. and Beer, A.). 169.
EBSWORTH, E. A. V. Volatile Silicon Compounds. 108.
ELLIOTT, R. M. (Editor). Advances in Mass Spectrometry, Vol. II. 48.
EVANS, JOHN W. (Editor). The Solar Corona. 47.
FIESER, LOUIS F., and FIESER, MARY. Topics in Organic Chemistry. 49.
FLORKIN, MARCEL and STOTZ, ELMER H. (Editors). Comprehensive Biochemistry, Vol. IX. 110.
FRIEDEL, J. and GUINIER, A. (Editors). Metallic Solid Solutions, a Symposium on their Electronic and Atomic Structure. 171.
GARRETT, S. D. Soil Fungi and Soil Fertility. 50.
GERSHENSON, DANIEL E. and GREENBERG, DANIEL A. (Editors). The Natural Philosopher. Vol. I. 172.
GOLD, V. (Editor). Advances in Physical Organic Chemistry, Vol. I. 168.
GRESHAM, G. A. and JENNINGS, A. R. An Introduction to Comparative Pathology, a Consideration of some Reactions of Human and Animal Tissues to Injurious Agents. 52.
HARRIS, R. J. C. (Editor). Biological Organization at the Cellular and Supercellular level. 109. Cellular Basis and Aetiology of Late Somatic Effects of Ionizing Radiation. 171.
HEGNAUER, R. Chemotaxonomie der Pflanzen, Vol. I. Thallophyten, Bryophyten, Pteridophyten und Gymnospermen. 50.
HERINGTON, E. F. G. Zone Melting of Organic Compounds. 109.

HODGMAN, C. D. (Editor-in-chief). Handbook of Mathematical Tables (supplement to Handbook of Chemistry and Physics). 47.

HUNTER, D. N. Inorganic Polymers. 168.

Institut Royal du Patrimoine Artistique, Bulletin VI, 1963. 171.

IVANOV, A. V. Pogonophora (translated from the Russian and edited by Carlisle, D. B., with additional material by Southward, Eve C.). 170.

JUCKER, ERNST (Editor). Progress in Drug Research, Vol. 5. 52.

KERKER, MILTON (Editor). Electromagnetic Scattering. 166.

KING, EDWARD L. How Chemical Reactions Occur: an Introduction to Chemical Kinetics and Reaction Mechanisms. 48.

KRISS, A. E. Marine Microbiology (Deep Sea) (translated from the Russian by Shewan, J. M. and Kabata, Z.) 51.

LEITMANN, GEORGE (Editor). Optimization Techniques with Applications to Aerospace Systems. 47.

LIPSCOMB, WILLIAM N. Boron Hydrides. 168.

MANDLEBERG, G. J. Topics in Modern Chemistry. 108.

MARR, J. W. S. The Natural History and Geography of the Antarctic Krill (*Euphausia Superba Dana*), Discovery Reports, Vol. XXXII. 111.

McDOWELL, C. A. (Editor). Mass Spectrometry. 167.

McILWAIN, HENRY. Chemical Exploration of the Brain (a Study of Cerebral Excitability and Ion Movement). 111.

MEITES, LOUIS (Editor). Handbook of Analytical Chemistry. 48.

MICHELSON, A. M. The Chemistry of Nucleosides and Nucleotides. 110.

MOELLER, THERALD, The Chemistry of the Lanthanides. 48.

NESMEYANOV, AN. N. Vapour Pressure of the Elements (translated and edited by Carasso, J. I.). 107.

NEWMAN, JAMES, R. Science and Sensibility, Vols. I and II. 107.

NIKOLSKY, G. V. The Ecology of Fishes (translated from the Russian by Birkett, L.). 51.

NISHIJIMA, K. Fundamental Particles. 165.

OSTEROTH, DIETER. Chemie und Technologie aliphatischer fluororganischer Verbindungen. 168.

PARR, N. L. (Editor). Laboratory Handbook. 165.

PERCIVAL, E. G. V. Structural Carbohydrate Chemistry (second edition, revised by Elizabeth Percival). 49.

PONTRYAGIN, L. S., BOLTYANSKII, V. G., GAMKRELIDZE, R. V., and MISHCHENKO, E. F. The Mathematical Theory of Optimal Processes (translated by Trirogoff, K. N.). 47.

PRIDHAM, J. B. (Editor). Enzyme Chemistry of Phenolic Compounds. 50.

RAINBOW, C. and ROSE, A. H. (Editors). Biochemistry of Industrial Micro-organisms. 169.

RATTRAY TAYLOR, G. The Science of Life: a Picture History of Biology. 112.

A Review of Space Research. 52.

RICHTER, N. B. The Nature of Comets (translated and revised by Beer, Arthur). 165.

RODD, E. H. (Editor). Chemistry of Carbon Compounds, Vol. V. Miscellaneous, General Index. 108.

ROSENBURG, H. M. Low Temperature Solid State Physics, Some Selected Topics. 166.

RUSSELL, F. S. (Editor). Advances in Marine Biology, Vol. I. 51.

SALISBURY, FRANK B. The Flowering Process. 170.

SANDLER, MAURICE and BOURNE, GEOFFREY H. (Editors). Atherosclerosis and its Origin. 170.

SCHMIDT, HELMUT and VON STACKELBERG, MARK. Modern Polarographic Methods (translated by Maddison, R. E. W.). 167.

SCHOFIELD, ROBERT E. The Lunar Society of Birmingham, a Social History of Provincial Science and Industry in Eighteenth-century England. 172.

SEGRÈ, E., FRIEDLANDER, G., and NOYES, H. P. (Editors). Annual Review of Nuclear Science, Vol. 13. 165.

SHINODA, K., NAKAGAWA, T., TAMAMUSHI, B., and ISEMURA, T. Colloidal Surfactants, some Physicochemical Properties. 108.

SHKLARSKY, D. O., CHENTZOV, N. N., and YAGLOM, I. M. The U.S.S.R. Olympiad Problem Book (revised and edited by Sussman, I., translated by Maykovich J.). 47.

SIENKO, M. J. and PLANE, R. A. Physical Inorganic Chemistry. 108.

SISSAKIAN, N. M. (General Editor). Proceedings of the Fifth International Congress of Biochemistry, Moscow, 1961, Vols. I–IX. 110.

SMEATON, W. A. Fourcroy: Chemist and Revolutionary. 52.

SMITH, WILSON (Editor). Mechanisms of Virus Infection. 49.

STANLEY, JOHN. The Essence of Biometry. 50.

STEPHEN, H. and STEPHEN, T. (Editors). Solubilities of Inorganic and Organic Compounds. Vol. I, Parts I and 2, Binary Systems. 167.

THOMSON, G. (Editor). Recent Advances in Conservation. 112.

Viruses, Nucleic Acids, and Cancer; a collection of papers presented at the 17th Annual Symposium of Fundamental Cancer Research 1963. 171.

WALEY, S. G. Mechanisms of Organic and Enzymic Reactions. 50.

WARING, H. Color Change Mechanisms of Cold-blooded Vertebrates. 110.

WELCHER, FRANK J. (Editor) Standard Methods of Chemical Analysis (sixth edition). Vols. IIA and IIB. Industrial and Natural Products and Noninstrumental Methods. 108.

WEST, P. W., MACDONALD, A. M. G., and WEST, T. S. (Editors). Analytical Chemistry 1962. 48.

WHISTLER, ROY L. and WOLFROM, M. L. (Editors). Methods in Carbohydrate Chemistry, Vol. I. Analysis and Preparation of Sugars. Vol. II. Reactions of Carbohydrates. Vol. III. (Edited by Roy L. Whistler only). Cellulose. 49.

WILSON, J. G. and WOUTHUYSEN, S. A. (Editors). Progress in Elementary Particle and Cosmic Ray Physics, Vol. VI. 107. Vol. VII. 165.

ZEPLER, E. E. and PUNNETT, S. W. (Editors). Electronic Devices and Networks. 52.

Subject index

The articles are arranged, with regard to subjects, in numerical order according to the Universal Decimal Classification (Brit. Std. 1000) An alphabetical key is to be found on page IV.

007 Information theory
007:621.391 The communication of information. Cherry, Colin. 13.

159.9 Psychology
159.929:591.51 The learning of discrimination by animals. Sutherland, N. S. 148.

37 Education
371.32–77 Books for students in developing countries (Editorial). 114.

5 Natural Sciences
51/53:92 Galileo Galilei (1564–1642) (Editorial). 2.

52 Astronomy
521:92 William Herschel and sidereal astronomy. Hoskin, Michael. 18.

53 Physics
53.082.8:543.544.25 Physical measurement by gas chromatography. Purnell, J. H. 142.
532.69:551.578.11 The collision, coalescence, and disruption of drops. Mason, B. J. 136.

537.312.62:538.66 Superconducting magnets. Kunzler, J. E. 115.
537.533:621.385.833 Field-ion microscopy. Brandon, D. G. 90.
538.6 Magnetohydrodynamics. Thompson, W. B. 73.
538.66:537.312 Superconducting magnets. Kunzler, J. E. 115.
539.166 Some applications of the Mössbauer effect. Moon, P. B. and O'Connor, D. A. 131.
539.4:548.526 Whiskers. Gordon, J. E. 8.

54 Chemistry

541.11 Modern thermochemistry. Mortimer, C. T. and Springall, H. D. 22.
541.183.5:553.066 Molecular sieves. Barrer, R. M. 122.
543.544.25:53.082.8 Physical measurement by gas chromatography. Purnell, J. H. 142.
546.02:92 Frederick Soddy and the concept of isotopes (Editorial). 54.
546.118 Plasma chemistry. Beguin, C. P., Kanaan, A. S., and Margrave, J. L. 55.
546.29 The chemistry of the noble gases. Bartlett, N. 3.
547.458:581.174 Structural plant polysaccharides. Preston, R. D. 153.
547.94:615.783 The relief of pain—the search for the ideal analgesic. Bentley, K. W. 97.
547.964.4:576.314:576.85 Teichoic acids and the bacterial cell wall. Baddiley, J. 33.

548 Crystallography

548.526:539.4 Whiskers. Gordon, J. E. 8.

551 Geophysics

551.324 Glaciology. Robin, G. de Q. 102.
551.345:624.139 Permafrost and related engineering problems. Brown, R. J. E. and Johnston, G. H. 66.
551.46(267) Geophysical studies in the Indian Ocean. Loncarevic, B. D. 43.
551.509.3 Weather forecasting by electronic calculation. Sutcliffe, R. C. 27.
551.578.11:532.69 The collision, coalescence, and disruption of drops. Mason, B. J. 136.
551.579 The movement of soil water. Childs, E. C. 81.

553 Mineralogy

553.066:541.183 Molecular sieves. Barrer, R. M. 122.

57 Biological Sciences. Biochemistry

576.314:578.6 Electron microscopy of the cell surface. Gray, E. G. 61.
576.314:576.85:547.964.4 Teichoic acids and the bacterial cell wall. Baddiley, J. 33.
576.85:547.964.4:576.314 *Idem.*
578.6:576.314 Electron microscopy of the cell surface. Gray, E. G. 61.

58 Botany

581.174:547.458 Structural plant polysaccharides. Preston, R. D. 153.
581.332:582.28 Germination of fungus spores. Gottlieb, David. 85.
582.28:581.332 *Idem.*

59 Zoology

591.51:159.929 The learning of discriminations by animals. Sutherland, N. S. 148.
591.53:595.772.4 Feeding habits of tsetse flies. Weitz, B. 38.
595.772.4:591.53 *Idem.*

61 Medical Sciences

615.783:547.94 The relief of pain—the search for the ideal analgesic. Bentley, K. W. 97.

62 Engineering Sciences

621.385.833:537.533 Field-ion microscopy. Brandon, D. G. 90.
621.391:007 The communication of information. Cherry, Colin. 13.
621.391:681.142 The language of machines. Bullard, Sir Edward. 160.
624.139:551.345 Permafrost and related engineering problems. Brown, R. J. E. and Johnston, G. H. 66.

68 Computers

681.142:621.391 The language of machines. Bullard, Sir Edward. 160.

9 Biography. History of Science

92:51/53 Galileo Galilei (1564–1642) (Editorial). 2.
92:521 William Herschel and sidereal astronomy. Hoskin, Michael. 18.
92:546.02 Frederick Soddy and the concept of isotopes (Editorial). 54.

Alphabetical key

Analgesics 615.783
Animal ecology 591.5
Astronomy 52
Bacterial cell wall 576.85
Biography 92
Botany 58
Cell surface structure 576.314
Chemistry 54
Computers, programming 681
Crystallography 548
Cytology 576.3
Drops, dynamic properties of 551.5
Education 37
Electrical engineering 621.3
Electron microscopy, cell surfaces 578.6
Field-ion microscopy 537.533
Fungus spore germination 582.28
Galileo 92
Gas chromatography 543.544
Geophysics 55
Germination of spores 581.332
Herschel, William 92
History of science 9
Indian Ocean, geophysics 551.46
Information theory 621.391
Language of machines 621.391
Learning, animals 591.51
Liquid drops, properties 532.69
Magnetohydrodynamics 538.6
Magnets, superconducting 538.66
Medical sciences 61
Meteorology 551.5
Mineralogy 553
Molecular sieves 541.183.5
Morphine structure 547.324
Mössbauer effect 539.166
Noble gases 546.29
Oceanography 551.46
Permafrost 551.345
Physics 53
Plant polysaccharides 581.174
Plasma chemistry 546.118
Polysaccharides 547.458
Psychology 159.9
Soddy, Frederick 92
Soil mechanics 624.139
Soil water movement 551.579
Superconductivity 537.312
Teichoic acids 547.964.4
Textbooks for students 371.32
Thermochemistry 541.11
Tsetse flies 595.772
Weather forecasting 551.509
Whiskers 548.526
Zeolites 553.066
Zoology 59

Made and printed in Great Britain at The Kynoch Press, Birmingham 6
for the publishers, Imperial Chemical Industries Limited, London S.W.1

077815

Some books received

(*Note.* Mention of a book on this page does not preclude subsequent review.)

GENERAL SCIENCE

The Dawn of a New Age, *by Eugene Rabinowitch. Pp. viii+332. The University of Chicago Press, Chicago and London. 1964. 52s. net.*

Science in its Context. A symposium with special reference to sixth-form studies. *Edited by John Brierley. Pp. vii+372. Heinemann Educational Books Ltd, London. 1964. 30s. net.*

MATHEMATICS

Introduction à la théorie des probabilités, *by B. V. Gnedenko and A. Ia. Khintchine (translated from the Russian into French by Marc Gilliard). Pp. x+157. Dunod, Paris. 1964. 15 Frs.*

Numerical Methods: 1. Iteration Programming and Algebraic Equations, *by B. Noble. Pp. xii+156. Oliver and Boyd Ltd., Edinburgh and London; Interscience Publishers Inc., New York. 1964. 10s. 6d. net.*

The Universal Encyclopedia of Mathematics *(translated and adapted from Meyers Rechenduden). Pp. 715. George Allen and Unwin Ltd, London. 1964. 42s. net.*

ASTRONOMY

Les hypothèses cosmogoniques. Théories des cycles cosmiques et des planètes jumelles, *by A. Dauvillier. Pp. 258. Masson et Cie., Paris. 1963. 29 Frs.*

The Planet Mercury, *by Werner Sandner (translated from the German by Alex Helm). Pp. 94. Faber and Faber Ltd, London. 1963. 21s. net.*

1964 Year Book of Astronomy. *Edited by J. G. Porter. Associate Editor: Patrick Moore. Pp. viii+215. Eyre and Spottiswoode, London. 1963. Cloth, 25s. net; paper, 16s. net.*

PHYSICS

The Electronic Theory of Catalysis on Semiconductors, *by F. F. Vol'kenshtein (translated from the Russian by N. G. Anderson, translation edited by E. J. H. Birch). Pp. vi+169. Pergamon Press, Oxford. 1963. 50s. net.*

Mandelstam Theory and Regge Poles. An Introduction for Experimentalists, *by R. Omnès and M. Froissart. Pp. xi+123. W. A. Benjamin Inc., New York. 1963. Cloth, $7.50; paper, $3.95.*

Nuclear Theory. Pairing Force Correlations and Collective Motion, *by A. M. Lane. Pp. x+250. W. A. Benjamin Inc., New York. 1964. Cloth, $8; paper, $4.95.*

Properties of the Thirty-two Point Groups, *by George F. Koster, John O. Dimmock, Robert G. Wheeler, and Hermann Statz. Pp. 104. The M.I.T. Press, Massachusetts Institute of Technology, Cambridge, Massachusetts. 1964. 57s. net.*

The Theory of Interacting Fermi Systems, *by P. Nozières (translated from the French by D. Hone). Pp. x+370. W. A. Benjamin Inc., New York. 1964. $13.50.*

CHEMISTRY

Acylation Reactions. Their Applications and Mechanisms, *by P. F. G. Praill. Pp. viii+162. The Commonwealth and International Library of Science, Technology, Engineering and Liberal Studies. Chemistry Division, Volume III. Pergamon Press, Oxford; The Macmillan Co., New York. 1963. 15s. net.*

Analytical Chemistry of Niobium and Tantalum, *by Ross W. Moshier. Pp. v+278. International Series of Monographs on Analytical Chemistry, Vol. XVI. Pergamon Press, Oxford. 1964. 90s. net.*

Atomic Structure and Chemical Bonding. A Non-mathematical Introduction, *by Fritz Seel (translated from the fourth German edition and revised by N. N. Greenwood and H. P. Stadler). Pp. vii+112. Methuen & Co. Ltd, London; John Wiley & Sons Inc., New York. 1963. 15s. net.*

Chemistry in Premixed Flames, *by C. P. Fenimore. Pp. ix+119. The International Encyclopedia of Physical Chemistry and Chemical Physics. Topic 19, Volume V. Pergamon Press, Oxford. 1964. 35s. net.*

The Inorganic Chemistry of Nitrogen, *by William L. Jolly. Pp. xi+124. W. A. Benjamin Inc., New York. 1964. $5.75.*

Inorganic Complexes, *by Chr. Klixbüll Jørgensen. Pp. v+220. Academic Press Inc. (London) Ltd, London. 1963. 42s. net.*

Kristallchemie, *by W. Kleber. Pp. 128. B. G. Teubner Verlagsgesellschaft, Leipzig. 1963. 9.30 DM.*

Mechanisms of Oxidation of Organic Compounds, *by W. A. Waters. Pp. 152. Methuen & Co. Ltd, London; John Wiley & Sons Inc., New York. 1964. 25s. net.*

Organic Polarographic Analysis, *by Petr Zuman. Pp. x+313. International Series of Monographs on Analytical Chemistry, Volume 12. Pergamon Press, Oxford. 1964. 50s. net.*

Oxidation Mechanisms: Applications to Organic Chemistry, *by Ross Stewart. Pp. xi+179. W. A. Benjamin Inc., New York. 1964. $7.50.*

Reaction Kinetics, *by Keith J. Laidler. Vol. I, pp. x+232. 21s. net. Vol. II, pp. x+136. 17s. 6d. net. The Commonwealth and International Library of Science, Technology, Engineering and Liberal Studies. Chemistry Division, Vols. I and II. Pergamon Press, Oxford. 1963.*

GEOPHYSICS

Climates of the Past. An Introduction to Paleoclimatology, *by Martin Schwarzbach (translated from the German and edited by Richard O. Muir). Pp. xii+328. D. Van Nostrand Co. Ltd, London. 1963. 63s. net.*

International Auroral Atlas. *Pp. 21+plates and figures. Published for the International Union of Geodesy and Geophysics by the Edinburgh University Press, Edinburgh. 1963. 45s. net.*

GEOLOGY

Fundamentals of Geology, *by Serge von Bubnoff (translated from the German and edited by W. T. Harry). Pp. xv+287. Oliver and Boyd Ltd, Edinburgh and London. 1964. 42s. net.*

Organic Geochemistry. *Edited by Irving A. Breger. Pp. x+658. International Series of Monographs on Earth Sciences, Volume XVI. Pergamon Press, Oxford. 1963. £7 net.*

Soil-geographical Zoning of the U.S.S.R. *(translated from the Russian by A. Gourevitch). Pp. 480. Distributed by the Oldbourne Press, London, for Israel Program for Scientific Translations, Jerusalem. 1964. £6 net.*

BIOLOGY

Cinemicrography in Cell Biology. *Edited by George G. Rose. Pp. xiv+500. Academic Press Inc., New York and London. 1963. 132s. net.*

The Ecology of North America, *by Victor E. Shelford. Pp. xxii+610. University of Illinois Press, Urbana, Illinois. 1963. $10.*

Germfree Life and Gnotobiology, *by Thomas D. Luckey. Pp. xii+512. Academic Press Inc., New York and London. 1963. 125s. net.*

International Review of Connective Tissue Research, *Vol. I. Edited by David H. Hall. Pp. xiii+401. Academic Press Inc., New York and London. 1963. $14.*

The Pathology of Development. A Study of Inherited Skeletal Disorders in Animals, *by Hans Grüneberg. Pp. xiv+309. Blackwell Scientific Publications, Oxford. 1963. 70s. net.*

Physiologie et pathologie du système lymphoïde, *by A. Policard. Pp. 193. Masson et Cie., Paris. 1963. 38 Frs.*

Survie et conservation biologique. *Edited by J. André Thomas. Pp. vii+478. Masson et Cie., Paris. 1963. 90 Frs.*

BIOCHEMISTRY

Biochemical Engineering, *by F. C. Webb. Pp. viii+743. D. Van Nostrand Co. Ltd, London. 1963. 120s. net.*

Einführung in die Praxis der Histochemie, *by Ludwig Spannhof. Pp. 150. Gustav Fischer Verlag, Jena. 1964. £1 10s. 4d. net.*

BOTANY

Physics of Plant Environment. *Edited by W. R. Van Wijk. Pp. xvi+382. North-Holland Publishing Co., Amsterdam. 1963. 80s. net.*

The Structure and Life of Bryophytes, *by E. V. Watson. Pp. 192. Hutchinson University Library, London. 1964. 15s. net.*

ZOOLOGY

Evolutionary and Genetic Biology of Primates, *Vol. I. Edited by John Buettner-Janusch. Pp. xiv+327. Academic Press Inc., New York and London. 1963. $12.*

The Genetics of the Silkworm, *by Yataro Tazima. Pp. xii+253. Academic Press Inc. (London) Ltd, London. 1964. 50s. net*

Insect Pathology. An Advanced Treatise, *Vol. II. Edited by Edward A. Steinhaus. Pp. xiv+689. Academic Press Inc., New York and London. 1963. 170s. net.*

Parasitism. An Introduction to Parasitology and Immunology for Students of Biology, Veterinary Science, and Medicine, *by J. F. A. Sprent. Pp. x+145. Baillière, Tindall and Cox Ltd, London. 22s. 6d. net.*

The Senses of Animals, *by L. Harrison Matthews and Maxwell Knight. Pp. 240. Museum Press Ltd, London. 1963. 27s. 6d. net.*

MEDICINE

Modern Developments in Audiology. *Edited by J. Jerger. Pp. xii+446. Academic Press Inc., New York and London. 1963. $12.*

TECHNOLOGY

Limestone as a Raw Material in Industry, *by F. P. Stowell. Pp. viii+55. Published under the auspices of Imperial Chemical Industries Ltd by Oxford University Press, London. 1963. 10s. 6d. net.*

Optimal Trajectories for Space Navigation, *by D. F. Lawden. Pp. viii+126. Butterworths, London. 1963. 21s. net.*

Progress in Ceramic Science, *Vol. III. Edited by J. E. Burke. Pp. vii+262. Pergamon Press, Oxford. 1963. 80s. net.*

HISTORY OF SCIENCE

Erasmus Darwin 1731–1802, *by Desmond King-Hele. Pp. vii+183. Macmillan & Co. Ltd, London; St. Martin's Press, New York. 1963 30s. net.*

Observationes, *Vol. I, by Felix Platter (translated from the Latin by Günther Goldschmidt, edited by Heinrich Buess). Pp. 198. Verlag Hans Huber, Berne and Stuttgart. 1963. Paper, 9.80 DM; cloth, 26.50 DM.*

Made and printed in Great Britain at The Kynoch Press, Birmingham 6
for the publishers, Imperial Chemical Industries Limited, Millbank, London S.W.1

Compiled
1999-2000